Speaker-Reasoner: Scaling Interaction Turns and Reasoning Patterns for Timestamped Speaker-Attributed ASR

Anonymous ACL submission

Abstract

Transcribing and understanding multi-speaker conversations requires speech recognition, speaker attribution, and timestamp localization. While speech LLMs excel at single-speaker tasks, multi-speaker scenarios remain challenging due to overlapping speech, backchannels, rapid turn-taking, and context window constraints. We propose Speaker-Reasoner, an end-to-end Speech LLM with agentic multi-turn temporal reasoning. Instead of single-pass inference, the model iteratively analyzes global audio structure, autonomously predicts temporal boundaries, and performs fine-grained segment analysis, jointly modeling speaker identity, gender, timestamps, and transcription. A speaker-aware cache further extends processing to audio exceeding the training context window. Trained with a three-stage progressive strategy, Speaker-Reasoner achieves consistent improvements over strong baselines across diverse multi-speaker benchmarks, including meeting scenarios, video domains, and multi-accent conversations, particularly in handling overlapping speech and complex turn-taking[1].

Figure 1: cpWER comparison across meeting, video, and multi-accent conversation domains. Speaker-Reasoner (4194h) is compared with Gemini-2.5-Pro and VibeVoice-ASR under the same evaluation protocol.

1 Introduction

In real-world multi-speaker conversational scenarios such as meetings and phone calls, comprehensive conversation understanding requires more than speech recognition alone. It demands the joint modeling of speaker attribution, fine-grained timestamp localization, and transcription (Gao et al., 2025; Yin et al., 2025). This task is essential for applications such as meeting transcription and intelligent assistants, yet it poses significant challenges including overlapping speech, backchannels, and rapid turn-taking, all of which degrade both speaker attribution and transcription accuracy.

Traditional approaches adopt cascaded frameworks that pipeline independent speaker diarization (SD) and ASR modules (Raj et al., 2021; Yu et al., 2022; Cornell et al., 2024; Böddeker et al., 2024). While modular in design, these systems suffer from error propagation between stages, difficulty coordinating outputs in the presence of overlapping speech, and the absence of joint optimization (Cornell et al., 2023; O'Shaughnessy, 2025; Yin et al., 2025). End-to-end approaches have been proposed to address these limitations, with serialized output training (SOT) (Kanda et al., 2020; Li et al., 2023) emerging as the dominant paradigm. However, SOT-based models rely on single-pass sequence-to-sequence inference and lack specialized mechanisms for complex phenomena such as simultaneous speech and interruptions.

Recent speech large language models (Speech LLMs) (Chu et al., 2024; KimiTeam et al., 2025;

[1]Code and models will be made publicly available

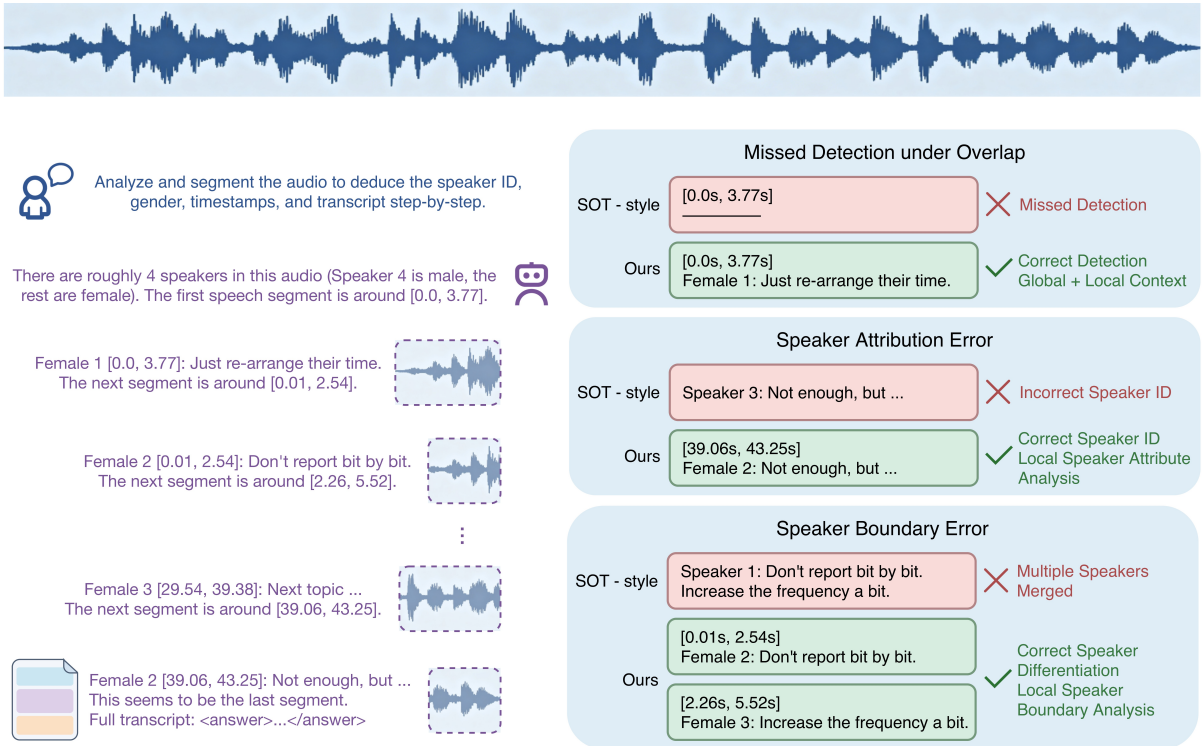

Figure 2: Illustration of the proposed reasoning process compared with SOT decoding. The left panel illustrates the iterative inference procedure of Speaker-Reasoner; the right panel provides a comparative case study.

Goel et al., 2025; StepFun Audio Team, 2025) such as MiMo-Audio (LLM-Core Xiaomi, 2025) and Qwen3-Omni (Xu et al., 2025) have demonstrated strong performance across a range of speech tasks, yet these models are primarily designed for single-speaker scenarios. Recent efforts (Yin et al., 2025; Peng et al., 2026; Huo et al., 2026; Shi et al., 2025; Meng et al., 2025b) have begun extending Speech LLMs to multi-speaker settings, but three key limitations persist. First, task coverage remains incomplete: full timestamped speaker-attributed ASR (TSA-ASR), which jointly models speaker identity, timestamps, and transcription, remains underexplored, and attributes such as speaker gender are often overlooked. Second, inference strategies remain simplistic, relying on single-pass SOT-style decoding without mechanisms tailored to overlapping speech or rapid turn-taking. Third, fixed context windows restrict processing to short segments, preventing the handling of long-form recordings common in real-world applications.

Recent visual reasoning advances (Lai et al., 2025; Meng et al., 2025a) demonstrate that complex reasoning problems can be addressed by scaling interaction turns and enriching reasoning patterns. Inspired by this, we observe that TSA-ASR inherently requires progressive reasoning along the temporal axis: the model must first understand the overall audio structure as global context, then progressively localize speaker boundaries, and finally perform fine-grained segment analysis. This global-to-local paradigm naturally aligns with multi-turn interaction mechanisms. Therefore, we propose Speaker-Reasoner, an end-to-end Speech LLM for TSA-ASR featuring agentic multi-turn reasoning on the temporal axis. Unlike existing methods performing single-turn sequence-to-sequence inference, Speaker-Reasoner employs iterative tool-based interactions: analyzing complete audio for global speaker distribution, autonomously predicting temporal boundaries, extracting and analyzing local slices, and reasoning about whether to continue or terminate.

The main contributions are as follows. (1) We propose *Speaker-Reasoner*, an end-to-end Speech LLM for TSA-ASR. To the best of our knowledge, this is the first work to introduce a multi-turn interactive mechanism to the TSA-ASR task, enabling iterative global-to-local temporal inference via boundary proposal and slice-level decoding. (2) We introduce a speaker-aware context cache to dynamically preserve cross-segment speaker consistency, extending this capability beyond the constraints of the fixed context window. (3) We develop a three-stage progressive training strategy to equip the model with multi-task speech under-

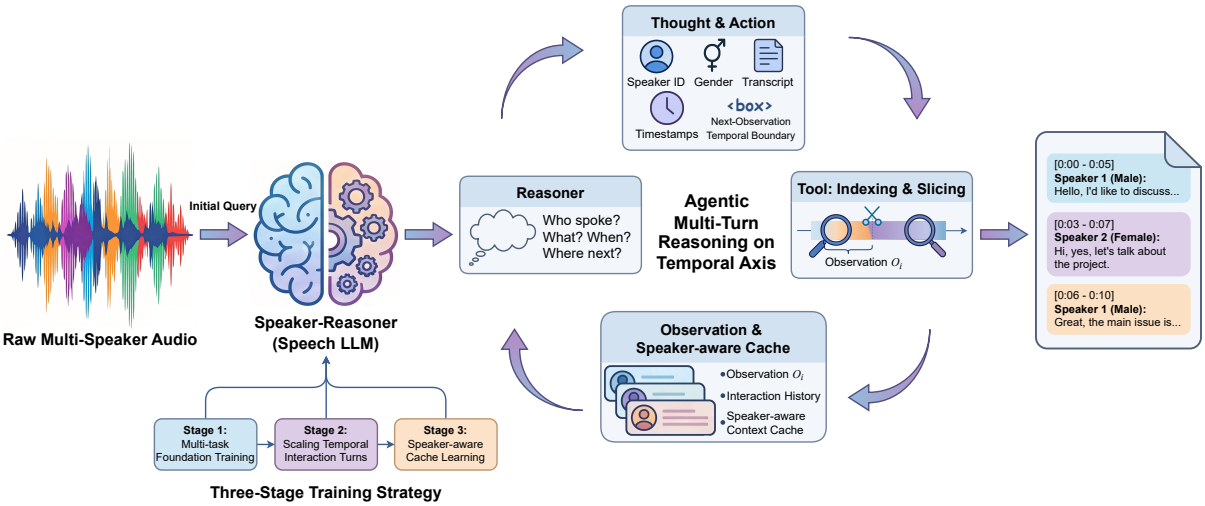

Figure 3: The overview of Speaker-Reasoner. The model employs an agentic multi-turn reasoning mechanism on the temporal axis, utilizing an indexing and slicing tool and a speaker-aware context cache to iteratively generate speaker identity, gender, timestamps, and transcription from raw multi-speaker audio.

standing, temporal interaction capability, and cache usage. (4) We scale the training data to 4,194 hours to thoroughly evaluate the model's performance under large-scale data regimes. Extensive experiments demonstrate that Speaker-Reasoner achieves highly competitive performance across a diverse range of primary benchmarks and domains, particularly excelling in challenging scenarios with overlapping speech and rapid turn-taking. To facilitate community research, our code and checkpoints will be made publicly available.

2 Method

Speaker-Reasoner addresses the TSA-ASR task in multi-speaker recordings. Given a raw recording, the model generates a chronologically ordered transcript, annotating each utterance with its speaker label, gender, and start and end timestamps.

A single-pass decoder is poorly suited for this setting. Overlapping speech and rapid turn-taking are difficult to resolve in a single step, while being constrained by fixed context windows. Therefore, we factorize inference into a global-to-local procedure progressing along the temporal axis through three coupled stages. First, a speaker inventory is estimated from a global recording view. The audio is then processed as a sequence of observations, representing contiguous time slices autonomously indexed and sequentially decoded. A speaker-aware context cache ensures consistent speaker identity. We term this iterative procedure multi-turn temporal interaction, where each turn corresponds to a single observation-decoding step. Figure 3 illustrates the overall architecture and inference flow.

2.1 Model Architecture

Speaker-Reasoner follows the standard audio-to-text Speech LLM architecture, consisting of an acoustic encoder, a modality projector, and an autoregressive LLM decoder. Unless otherwise stated, all components are initialized from Qwen3-Omni (Xu et al., 2025).

Acoustic encoder We adopt the Audio Transformer (AuT) encoder from Qwen3-Omni. A Conv2D module downsamples the input Fbank features by a factor of 8, yielding an acoustic feature sequence at a frame rate of 12.5 Hz, which is subsequently processed by stacked Transformer blocks.

Projector A two-layer MLP projects the acoustic features into the LLM embedding space, enabling acoustic and text tokens to share a unified representation.

LLM decoder We adopt the Thinker module from Qwen3-Omni, a Mixture-of-Experts (MoE) Transformer that activates a sparse subset of experts per token, maintaining high model capacity while ensuring computational efficiency. The decoder receives the concatenation of the projected audio features $\mathbf{z}$, text instructions, and previously generated tokens to generate responses autoregressively.

2.2 Iterative Temporal Interaction

The model traverses the recording incrementally along the temporal axis. At turn i, it takes as input the current observation O_i, the decoding his-

tory $\mathcal{H}_{i-1}$ accumulated over the previous turns, and the speaker-aware cache $\mathcal{C}_{i-1}$ (Section 2.3). It then outputs the speaker identity, gender, timestamps, and transcription for every utterance within O_i, along with the temporal boundary for the next observation O_{i+1}. A turn denotes one such decoding step on a single audio slice, and interaction refers to the cumulative dependency across turns established by $\mathcal{H}_{i-1}$ and $\mathcal{C}_{i-1}$.

Observation definition. An observation is a contiguous time slice of the original audio, $O_i = x[t_i^{\text{st}} : t_i^{\text{ed}}]$, with $0 \leq t_1^{\text{st}} < t_2^{\text{st}} < \cdots < t_{N+1}^{\text{st}} \leq T$. The model is conditioned on the triple $(O_i, \mathcal{H}_{i-1}, \mathcal{C}_{i-1})$ at each turn.

Boundary construction and supervision. Observation boundaries are derived from speaker-level utterance segments, compelling the model to explicitly handle turn-taking and partial overlaps. Based on the ground-truth utterance segmentation of the recording, we group utterances into observations as follows. For two utterances A and B with durations d_A and d_B, and an overlap duration of d_{AB}, we consider them to be heavily overlapping if

$$\frac{d_{AB}}{d_A} \geq \tau \quad \text{or} \quad \frac{d_{AB}}{d_B} \geq \tau, \tag{1}$$

where $\tau = 0.8$ in our experiments. Heavily overlapping utterances are merged into a single observation and decoded jointly. This joint decoding is necessary for backchannels and simultaneous speech, where no clear separating boundary exists. Each of the remaining utterances constitutes an independent observation, and the subsequent observation O_{i+1} is defined as the nearest unit following O_i. During inference, the model generates the start and end points of O_{i+1} as discrete boundary tokens, eliminating the requirement for an external voice activity detector or fixed-stride chunking.

Interaction protocol. The first turn is dedicated to global analysis, during which the model processes the entire audio recording and outputs a summary of the speaker count and gender distribution. This summary is appended to $\mathcal{H}_0$, serving as a global anchor for subsequent local decoding. From the second turn onward, the decoding of each observation O_i is conditioned on $\mathcal{H}_{i-1}$ and $\mathcal{C}_{i-1}$ as described above. The interaction terminates when the model encloses the consolidated transcript within `<answer>` and `</answer>` tags, yielding a chronologically sorted speaker-attributed transcript.

2.3 Speaker-aware Context Cache

Long-form audio often exceeds the training context window, causing speaker identity drift across chunks. We introduce a speaker-aware cache that supplies speaker-specific acoustic references from previously processed content.

Cache entry representation. Each cache entry is a tuple: $c = \big(s, \tilde{x}[t_{\text{st}} : t_{\text{ed}}], \tilde{y}\big)$, where s is the speaker label, $\tilde{x}$ is an audio slice, and $\tilde{y}$ is the corresponding transcript. Cache entries are prepended to the prompt as structured context.

Cache training simulation. To expose the model to cache usage during training, we employ a random sampling strategy. For each training instance, we randomly select a subset of speakers from the same session and retrieve 1–5 segments per speaker from non-overlapping earlier audio regions, which are prepended as cache context. Speaker labels are reassigned by order of first appearance in the augmented input sequence, encouraging permutation-invariant speaker assignment and ensuring label continuity between the cached history and the current observation.

Cache selection at inference. During inference, we maintain an observation buffer $B = \{O_1, \ldots, O_n\}$ of previously processed segments. For each speaker $s \in S$, candidate segments are scored jointly by duration and recency: $\phi = d \cdot \big(1 + \alpha \cdot \frac{i}{n}\big)$, where d is the segment duration, i is its position in the buffer, and α controls the recency weight. We retain the top-k segments per speaker as cache entries (Algorithm 1). These entries provide speaker-specific acoustic references that stabilize identity assignment for both recurring and newly emerging speakers.

2.4 Training Objective and Curriculum

We adopt a three-stage training curriculum that progressively builds multi-task perception, temporal interaction, and long-context robustness within the model.

Stage 1: Multi-task foundation. We train the model with teacher forcing to generate the complete structured output sequence for each recording, which includes speaker labels, gender, timestamps, and transcripts. The training objective is a standard next-token prediction loss:

$$\mathcal{L}_{\text{LM}} = -\sum_t \log p_\theta(w_t \mid w_{<t}, \mathbf{z}), \tag{2}$$

where w_t denotes the t-th output token. Prepend-

Algorithm 1 Speaker-aware Cache Selection

Require: Observation buffer $B = \{O_1, \ldots, O_n\}$, speaker set S, recency weight α, cache size k

Ensure: Selected cache entries C

```
Initialize empty candidate list L_s for each s ∈ S
for i = 1 to n do
    ρ ← i/n                                   ▷ recency ratio
    for each segment (s, t_st, t_ed, text) in O_i do
        if s ∈ S then
            d ← t_ed − t_st
            φ ← d · (1 + α · ρ)
            Append (O_i, φ) to L_s
for each speaker s ∈ S do
    Sort L_s by φ in descending order
    C[s] ← top-k entries of L_s
return C
```

ing a global summary header that lists the speaker count and gender distribution encourages global-to-local processing awareness.

Stage 2: Temporal interaction learning. In this stage, the model learns to process each observation O_i conditioned on the prior interaction history. Along with the language modeling loss, we supervise predicted boundary tokens:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda \, \mathcal{L}_{\mathrm{bd}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{bd}}$ is a cross-entropy loss over discretized boundary bins and λ is a balancing weight.

Stage 3: Speaker-aware cache learning. To support cache-conditioned decoding, we prepend sampled historical cache entries during training to mirror the prompting mechanism used at inference. This stage improves speaker consistency across long-form audio by exposing the model to cross-chunk speaker references and enforcing label continuity.

3 Experiments

3.1 Implementation Details

We initialize Speaker-Reasoner from Qwen3-Omni, a 30B-parameter multimodal LLM with an MoE architecture that activates 3B parameters per forward pass. To investigate the impact of model scale on performance, we also develop a 7B variant initialized from Qwen2.5-Omni, maintaining identical training configurations. Training is conducted using the MS-Swift framework (Zhao et al., 2025) with a Megatron-LM backend on 8 NVIDIA A100 GPUs. We apply LoRA with rank 8 and scaling factor 32 to all linear layers across the audio encoder, projector, and language model. We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 1e-5, cosine schedule, 0.05 warmup fraction, and minimum learning rate of 1e-6. The three training stages are executed sequentially, with each stage initialized from the checkpoint of the previous stage.

3.2 Datasets

Our primary public meeting benchmarks are AISHELL-4 (Fu et al., 2021) and AliMeeting (Liang et al., 2023). AISHELL-4 contains 211 Mandarin meeting sessions with 4–8 speakers and roughly 120 hours of speech. AliMeeting contains 424 Mandarin meeting sessions with 2–4 participants and roughly 120 hours of speech. Both benchmarks include realistic meeting phenomena such as reverberation, natural pauses, overlap, and turn-taking, and provide speaker labels, timestamps, and manual transcriptions.

For the scaled setting, we train on 4,194 hours spanning meetings, conversations, and video audio. The corpus includes approximately 1,200 hours of Chinese meeting data from AISHELL-4, AliMeeting, and similar internal meetings; approximately 1,000 hours of multi-dialect Chinese conversational speech; English meeting audio from AMI-SDM1 (Carletta et al., 2005); multi-accent English conversation data from MLC-SLM (Mu et al., 2025) Tracks 1 and 2; and approximately 1,200 hours of internal video audio from TV series, movies, and animations.

Following the segmentation protocol in prior work (Yin et al., 2025), we use only the first microphone channel and split the original long-form recordings into segments of 40 to 50 seconds based on voice activity detection boundaries for both training and evaluation.

Evaluation is organized by application domain. Meeting evaluation uses AISHELL4-Eval, Alimeeting-Far, and AMI-eval SDM1. Multi-accent robustness is measured on MLC-SLM Track 1 and Track 2 English subsets. Video evaluation uses internal Chinese and English test sets with 2–6 speakers per recording and challenging background noise or music.

Training proceeds in stages. The first stage uses

Table 1: Performance comparison on multiple benchmarks. The evaluation is categorized into three domains: Video, Meeting, and Conversation (MLC-SLM). **Bold** indicates the best result in each dataset.

Domain	Dataset	Models (Top to Bottom): Gemini-2.5-Pro, VibeVoice-ASR, Ours			
		DER↓	WER↓	cpWER↓	Δcp↓
Video Domain	**Video-Internal (ZH&EN)**				
	Overall (ZH & EN), 2–6 spk, TV series / short drama / anime, noisy background	74.05	22.47	44.13	21.66
		47.18	16.45	58.60	42.15
		15.33	**6.27**	**24.43**	**18.16**
	Video-Internal (ZH)				
	Mandarin, 2–6 spk, TV series / short drama / anime, noisy background	69.35	18.28	40.97	22.69
		47.65	17.70	62.06	44.36
		16.68	**6.50**	**25.81**	**19.31**
	Video-Internal (EN)				
	English, 2–6 spk, TV series / short drama / anime, noisy background	100.95	55.40	68.82	13.42
		44.62	7.11	32.65	25.54
		7.58	**4.42**	**16.31**	**11.89**
Meeting Domain	**AISHELL4-Eval (ZH)**				
	Mandarin, 4–8 spk, Real conference, 8-ch microphone array	36.07	19.81	25.11	5.30
		8.94	22.19	26.16	3.97
		3.38	**7.13**	**8.14**	**1.01**
	Alimeeting-Far (ZH)				
	Mandarin, 2–4 spk, Real meeting, 8-ch far-field array	56.39	30.16	39.29	9.13
		19.62	34.31	39.92	5.61
		6.70	**19.72**	**19.92**	**0.20**
	AMI-SDM (EN)				
	English, ∼4 spk, Scenario meeting, Single Distant Microphone	50.28	31.66	39.98	8.32
		21.00	30.53	35.86	5.33
		13.56	**23.29**	**25.16**	**1.87**
Conversation	**MLC-SLM-Track1 (EN)**				
	English subset, multi-accent, 2 spk, Spontaneous dialogue	42.33	36.87	41.88	5.01
		6.27	10.30	13.45	3.15
		4.76	**9.17**	**11.74**	**2.57**
	MLC-SLM-Track2 (EN)				
	English subset, multi-accent, 2 spk, Spontaneous dialogue	46.19	26.73	32.19	5.46
		3.14	**7.97**	**11.38**	3.41
		4.35	8.54	11.76	**3.22**

the complete 4,194-hour corpus for foundation pre-training, exposing the model to diverse speech patterns and acoustic conditions. Subsequent supervised fine-tuning stages employ a carefully filtered 2,000-hour subset focused on the highest-quality annotations and the most challenging scenarios. This strategy enhances model robustness across diverse languages, accents, and interaction patterns, while maintaining computational efficiency during fine-tuning.

3.3 Evaluation Metrics

We evaluate speaker diarization using Diarization Error Rate (DER), which measures the fraction of time incorrectly attributed to speakers, encompassing missed speech, false alarms, and speaker confusion errors. For transcription quality, we report Character Error Rate (CER) for Chinese test sets and Word Error Rate (WER) for English test sets. To jointly assess transcription accuracy and speaker attribution, we adopt the concatenated minimum-permutation error rate (cpCER for Chinese and cpWER for English), which concatenates all utterances per speaker, finds the optimal speaker permutation, and then computes the respective error rate. We additionally report Δcp, defined as the difference between the concatenated error rate (cpCER or cpWER) and the standard error rate (CER or WER), which isolates the error contribution from speaker assignment by factoring out intrinsic transcription errors.

Table 2: Comparison of speaker diarization and recognition performance on AISHELL4-Eval and Alimeeting-Far test sets. The proposed models also serve as a stage-wise ablation across training stages. **Bold** denotes the best result in each column.

Model	AISHELL4-Eval				Alimeeting-Far			
	DER↓	CER↓	cpCER↓	Δcp↓	DER↓	CER↓	cpCER↓	Δcp↓
Cascade								
Pyannote3.1 + Paraformer	8.10	19.18	26.24	7.06	19.13	30.15	45.39	15.24
End-to-End								
Gemini-2.5-Pro†	36.07	19.81	25.11	5.30	56.39	30.16	39.29	9.13
Qwen3-Omni-30B-A3B-Instruct	32.42	14.46	22.22	7.76	37.15	25.40	36.28	10.88
Qwen2.5-Omni-7B	85.68	33.37	60.45	27.08	91.77	38.13	73.38	35.25
SpeakerLM (212.25h)	–	17.75	26.14	8.39	–	18.63	32.22	13.59
SpeakerLM (7638.95h)	–	17.17	18.37	1.20	–	**13.97**	**16.05**	2.08
VibeVoice-ASR	10.88	22.30	26.30	4.00	20.70	34.67	40.54	5.87
TagSpeech-Alimeeting	37.51	35.70	53.44	17.74	52.46	47.11	68.74	21.63
Ours (Stage-wise Ablation)								
Qwen3-Omni + SOT sft	–	17.65	19.59	1.94	–	24.24	26.03	1.79
Speaker-Reasoner Base (Stage 1)	6.24	14.04	16.54	2.50	8.96	21.16	22.64	1.48
Speaker-Reasoner Multi-turn (Stage 2)	5.19	13.83	14.93	1.10	7.47	20.34	20.29	-0.05
Speaker-Reasoner Multi-turn w/ SAC (Stage 3)	5.26	13.83	14.73	**0.90**	7.34	20.57	20.43	**-0.14**
Speaker-Reasoner Multi-turn 4194h	**3.38**	**7.13**	**8.14**	1.01	**6.70**	19.72	19.92	0.20
Speaker-Reasoner Base 7B	12.00	15.65	25.60	9.95	18.43	24.97	38.12	13.15
Speaker-Reasoner Multi-turn 7B	9.38	15.31	22.91	7.60	15.56	24.33	34.81	10.48

† Closed-source model. DER unavailable for SpeakerLM and SOT-based models due to incompatible output formats.

3.4 Performance Comparison with Baselines

To comprehensively evaluate the proposed approach, we benchmark our large-scale Speaker-Reasoner (4194h) against open-source VibeVoice-ASR and closed-source Gemini-2.5-Pro across multiple domains. Table 1 shows Speaker-Reasoner (4194h) achieving SOTA performance across most primary metrics. On internal video and AMI-SDM evaluation sets, the model consistently yields lower error rates than the baselines. For multi-accent conversation, it demonstrates robustness across diverse accents; although its performance on the Track 2 subset slightly trails VibeVoice-ASR, the closely comparable error rates confirm its TSA-ASR capabilities in complex acoustic environments.

Table 2 presents a detailed comparison on the AISHELL4-Eval and Alimeeting-Far test sets against cascaded frameworks, closed-source systems, and other end-to-end models. On AISHELL4-Eval, the Speaker-Reasoner Multi-turn w/ SAC model yields 5.26% DER and 14.73% cpCER. On Alimeeting-Far, the model obtains 7.34% DER and 20.43% cpCER. When expanding the training corpus to 4,194 hours (Speaker-Reasoner Multi-turn 4194h), the model further reduces the DER to 3.38% and cpCER to 8.14% on AISHELL4-Eval, alongside 6.70% DER and 19.92% cpCER on Alimeeting-Far. These metrics indicate consistent improvements over the baselines under identical test conditions.

Notably, our best model achieves a Δcp of -0.14% on Alimeeting-Far. While Δcp is typically positive due to speaker attribution errors, a negative value indicates cpCER is slightly lower than standard CER. This occurs because concatenated minimum-permutation alignment in the cpCER calculation compensates for minor transcription inaccuracies, such as boundary-related deletions or insertions, provided the model exhibits high precision in speaker tracking and reasoning.

Finally, we examine the impact of training paradigms and model scales. Compared to the conventional single-pass sequence-to-sequence framework (Qwen3-Omni + SOT sft), our iterative multi-turn interaction mechanism achieves lower error rates under equivalent scaling conditions. To verify efficiency in resource-constrained settings, we evaluate the 7B variants initialized from Qwen2.5-Omni. Speaker-Reasoner Multi-turn 7B outperforms the zero-shot Qwen2.5-Omni-7B baseline, achieving 22.91% cpCER on AISHELL4-Eval and 34.81% on Alimeeting-Far. This performance exceeds specialized baselines trained on larger datasets, demonstrating the proposed reasoning

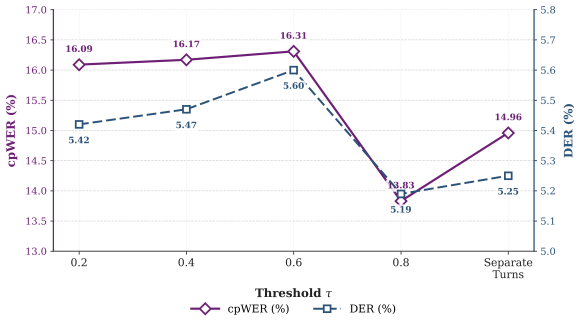

Figure 4: Ablation study on the observation boundary threshold τ.

framework's effectiveness even at smaller parameter scales.

3.5 Ablation on Training Stages

Table 2 illustrates performance evolution across the three training stages of Speaker-Reasoner. Transitioning from the base model to the multi-turn interaction stage yields substantial improvements, simultaneously reducing DER, cpCER, and Δcp across both evaluation sets. This confirms the agentic reasoning paradigm effectively captures complex temporal dependencies and speaker transitions. Furthermore, the speaker-aware cache (SAC) introduced in the final stage dynamically preserves cross-segment speaker consistency, extending this capability beyond fixed context window constraints. Stable DER and cpCER metrics upon SAC integration demonstrate that the cache successfully extends the model's capabilities without compromising its diarization or recognition accuracy.

3.6 Ablation on Observation Boundary Threshold

We investigate the impact of the overlap threshold τ, used for training observation boundary construction (Section 2.2), on overall performance. We evaluate $\tau \in \{0.2, 0.4, 0.6, 0.8\}$ alongside a Separate Turns baseline, where every utterance forms an independent interaction turn regardless of temporal overlap. Figure 4 shows lower thresholds ($\tau \leq 0.6$) yield higher error rates, with cpWER exceeding 16.00% and DER remaining above 5.40%. A lower τ aggressively merges adjacent segments with only partial overlaps, overcrowding observation windows and increasing the complexity of local transcription and speaker tracking. Conversely, the Separate Turns approach entirely avoids segment merging. While achieving moderate results (cpWER 14.96%, DER 5.25%), it forcibly splits highly concurrent speech events, preventing joint modeling of tightly coupled acoustic contexts for simultaneous utterances like backchannels. Optimal performance occurs at $\tau = 0.8$, achieving the lowest cpWER of 13.83% and DER of 5.19%. This threshold effectively isolates sequential turn-taking while preserving joint processing of heavily overlapping speech, striking an optimal balance for multi-turn temporal reasoning.

Table 3: Gender prediction accuracy (ACC) and speaker count accuracy (SCA) on AISHELL4-Eval. SCA measures whether the predicted number of speakers matches the ground truth.

Model	ACC	SCA
Gemini-2.5-Pro	94.80	67.03
Qwen3-Omni	97.12	60.49
Speaker-Reasoner Multi-turn	**96.80**	**69.03**

3.7 Evaluation of Speaker Attributes

Table 3 evaluates gender prediction accuracy (ACC) and speaker count accuracy (SCA) on AISHELL4-Eval. Speaker-Reasoner Multi-turn achieves a gender accuracy of 96.80% and an SCA of 69.03%, outperforming both Gemini-2.5-Pro and the zero-shot Qwen3-Omni baseline on SCA. The improved speaker count accuracy indicates that the iterative global-to-local reasoning paradigm enables the model to more reliably distinguish individual speakers within complex multi-speaker audio, leading to better global speaker composition estimation.

4 Conclusion

We present Speaker-Reasoner, an end-to-end Speech LLM for TSA-ASR. Optimized via a three-stage progressive strategy, its agentic multi-turn reasoning shifts inference to an iterative global-to-local paradigm. This enables the model to autonomously resolve overlapping speech and rapid turn-taking through boundary proposal and slice-level decoding, while a speaker-aware context cache ensures identity consistency beyond fixed context windows. Scaled to 4,194 hours, it achieves highly competitive performance across diverse meeting, video, and multi-accent domains. This demonstrates the potential of scaling interaction turns and reasoning patterns to enhance comprehensive conversational understanding in Speech LLMs.

5 Limitations

This work has several limitations. First, limited access to large-scale, diverse open-source data necessitates the use of internal corpora for training and evaluation, which may constrain direct reproducibility despite the release of comprehensive statistics and scripts. Second, the system relies on predicted boundary tokens, where suboptimal predictions may compound across iterative turns. Furthermore, gender prediction is bounded by the label schema of the evaluated corpora, potentially failing to capture non-binary or socially complex gender identities. Finally, the multi-turn inference paradigm introduces higher decoding costs than single-pass SOT; future research should therefore quantify the latency-accuracy trade-offs required for streaming or real-time deployment.

6 Ethical Considerations

This work uses existing benchmarks, pretrained foundation models, and internal audio datasets for research and evaluation only. We do not actively collect new human-subject conversational data or recruit human annotators for this study, and we will not redistribute proprietary third-party benchmark data. Since timestamped speaker-attributed ASR systems may be deployed in sensitive speech or speaker-related settings, outputs from Speaker-Reasoner should be treated as decision support rather than ground truth, especially for personal or legally consequential audio. Additionally, because the model's gender prediction follows the historical labels available in the evaluated corpora, it may not capture non-binary or socially complex gender identities, requiring caution to avoid representational bias in real-world deployment. Finally, the authors utilized an AI assistant to assist in the linguistic refinement and polishing of this manuscript.

References

Christoph Böddeker, Aswin Shanmugam Subramanian, Gordon Wichern, Reinhold Haeb-Umbach, and Jonathan Le Roux. 2024. TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:1185–1197.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried M. Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proc. MLMI*, pages 28–39.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

Samuele Cornell, Jee-Weon Jung, Shinji Watanabe, and Stefano Squartini. 2024. One model to rule them all ? towards end-to-end joint speaker diarization and speech recognition. In *Proc. ICASSP*, pages 11856–11860. IEEE.

Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola García, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur. 2023. The chime-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios. *CoRR*, abs/2306.13734.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 2021. AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In *Proc. Interspeech*, pages 3665–3669.

Ming Gao, Shilong Wu, Hang Chen, Jun Du, Chin-Hui Lee, Shinji Watanabe, Jingdong Chen, Sabato Marco Siniscalchi, and Odette Scharenborg. 2025. The multimodal information based speech processing (MISP) 2025 challenge: Audio-visual diarization and recognition. In *Proc. Interspeech*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *CoRR*, abs/2507.08128.

Mingyue Huo, Yiwen Shao, and Yuheng Zhang. 2026. Tagspeech: End-to-end multi-speaker ASR and diarization with fine-grained temporal grounding. *CoRR*, abs/2601.06896.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2020. Serialized output training for end-to-end overlapped speech recognition. In *Proc. Interspeech*, pages 2797–2801.

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. *CoRR*, abs/2504.18425.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 2025. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *CoRR*, abs/2509.07969.

Chenda Li, Yao Qian, Zhuo Chen, Naoyuki Kanda, Dongmei Wang, Takuya Yoshioka, Yanmin Qian, and Michael Zeng. 2023. Adapting multi-lingual ASR models for handling multiple talkers. In *Proc. Interspeech*, pages 1314–1318.

Yuhao Liang, Mohan Shi, Fan Yu, Yangze Li, Shiliang Zhang, Zhihao Du, Qian Chen, Lei Xie, Yanmin Qian, Jian Wu, Zhuo Chen, Kong Aik Lee, Zhijie Yan, and Hui Bu. 2023. The second multi-channel multi-party meeting transcription challenge (m2met 2.0): A benchmark for speaker-attributed ASR. In *Proc. ASRU*, pages 1–8. IEEE.

LLM-Core Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners. *CoRR*, abs/2512.23808.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. 2025a. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *CoRR*, abs/2510.20579.

Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2025b. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. In *Proc. ICASSP*, pages 1–5. IEEE.

Bingshen Mu, Pengcheng Guo, Zhaokai Sun, Shuai Wang, Hexin Liu, Mingchen Shao, Lei Xie, Eng Siong Chng, Longshuai Xiao, Qiangze Feng, and Daliang Wang. 2025. Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods. *CoRR*, abs/2509.13785.

Douglas O'Shaughnessy. 2025. Speaker diarization: A review of objectives and methods. *Applied Sciences*, 15(4).

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, Yutao Sun, Hangbo Bao, Weijiang Xu, Yi Zhu, Zehua Wang, Ting Song, Yan Xia, Zewen Chi, Shaohan Huang, and 5 others. 2026. VIBEVOICE-ASR technical report. *CoRR*, abs/2601.18184.

Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, Zili Huang, Maokui He, Shinji Watanabe, Jun Du, Takuya Yoshioka, Yi Luo, Naoyuki Kanda, Jinyu Li, Scott Wisdom, and John R. Hershey. 2021. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In *Proc. SLT*, pages 897–904. IEEE.

Mohan Shi, Xiong Xiao, Ruchao Fan, Shaoshi Ling, and Jinyu Li. 2025. Train short, infer long: Speech-llm enables zero-shot streamable joint asr and diarization on long audio. *CoRR*, abs/2511.16046.

StepFun Audio Team. 2025. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

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, and 19 others. 2025. Qwen3-omni technical report. *CoRR*, abs/2509.17765.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2025. Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. *CoRR*, abs/2508.06372.

Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, Yanmin Qian, Kong Aik Lee, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022. Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge. In *Proc. ICASSP*, pages 9156–9160. IEEE.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proc. AAAI*, pages 29733–29735.