Thinking-while-speaking: A Controlled, Interleaved Reasoning Method for Real-Time Speech Generation

Anonymous ACL submission

Abstract

The *thinking-while-speaking* paradigm aims to make AI communication more human. A key challenge is maintaining fluent speech while performing deep reasoning. Our method, **InterRS**, tackles this by inserting reasoning steps only during natural speech generation. This requires high-quality data where reasoning and speech are precisely aligned, and the length ratio are under controlled. We introduce a novel pipeline to generate such seamlessly interleaved audio data. To train our model, we combine interleaved SFT with refined data and reinforcement learning with two new rewards: a *TA-Balance Reward* to manage timing and thinking-answer ratio, and a *Linguistic Quality Reward* to refine expression. Experiments show our approach achieves **13%** better performance on mathmatical and logic benchmarks while **generating instant response** like a spoken-language instruct model which outputs fast CoT response. Furthermore, our method generates more natural and fluent answers than prior methods.

1 Introduction

Spoken language models (SLMs) are rapidly advancing (Lakhotia et al., 2021; Rubenstein et al., 2023; Borsos et al., 2023; Zhang et al., 2023) . This progress has made real-time interaction a key research focus, aiming for more natural and human-like communication (Chu et al., 2024; Défossez et al., 2024; Wang et al., 2026). When solving complex logical tasks, humans often exhibit a *thinking-while-speaking* pattern (Sacks et al., 1974). They continuously perform internal reasoning while expressing their thoughts (Ma et al., 2025).

As a comparable early attempt to incorporate reasoning capabilities into SLMs, researchers are adapting the chain-of-thought (CoT) paradigm from text models for speech (Wei et al., 2022; Ma et al., 2025; Du et al., 2024a; Kojima et al., 2022b). By leveraging reinforcement learning, the model's capacity for reasoning can be further improved. However, the traditional *thinking-before-speaking* approach faces major challenges in speech interaction (Défossez et al., 2024). The length of a reasoning chain is often unpredictable. Requiring a model to finish all reasoning before speaking causes unacceptably high first token latency. This delay breaks the real-time flow of conversation.

To improve the user experience, recent works explore an interleaved *thinking-while-speaking* paradigm (Xie et al., 2025; Yu et al., 2024b). The goal is to perform supplementary thinking during speech generation. Yet, current interleaved methods have two key shortcomings. First, the alignment between thinking and speaking content can be rigid. This makes it hard to balance reasoning depth with verbal fluency. Second, training often lacks effective control over the thinking-answer ratio. This can lead to overly long replies and introduce undesired pause in speech generation.

To solve these problems, we propose a new reasoning paradigm named **InterRS** for speech models, as show in Figure 1. It leverages the temporal gaps in speech generation for supplementary reasoning. This allows for significantly more accurate and fluent responses while maintaining real-time performance. Our main contributions are threefold:

- **A framework for fluent, controllable interleaved reasoning data.** We designed a fine-grained data alignment mechanism. It segments and associates speaking content with the underlying thinking process based on semantic logic, allowing control over the thinking ratio of data. It ensures natural transitions and semantic consistency in the interleaved speech.
- **A training strategy with two novel rewards.** The interleaved SFT phase is designed to endow the model with the foundational capa-

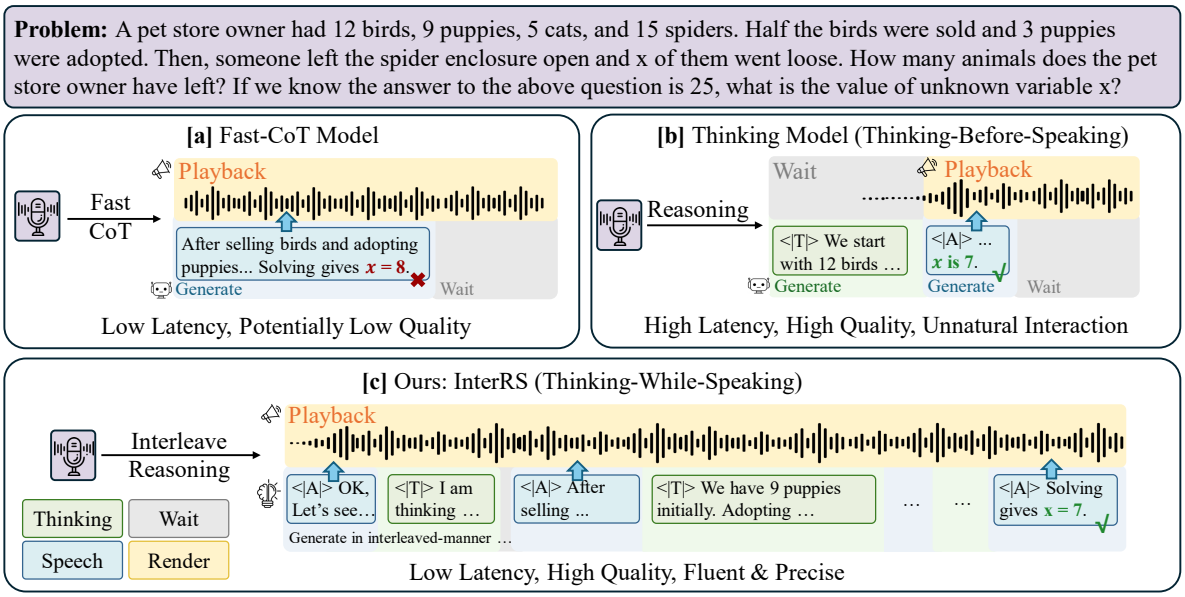

Figure 1: Our framework, InterRS, integrates reasoning steps within temporal gaps in speech generation, enabling seamless thought-speech alignment. It advances beyond traditional interleaved methods by jointly optimizing fluency and accuracy through two novel reward functions, achieving improved performance on reasoning tasks and perceived response naturalness.

bility for controlled output length and fluent articulation. For the post-training phase, we introduce two reward functions. The TA-Balance Reward dynamically regulates the interleave reasoning length to prevent overthinking. The Linguistic Quality Reward is optimized to maintain fluent and stable speech flow during reasoning interleaving.

- **Strong performance on logical & mathematical reasoning.** Our method shows significant performance gains (+13% compared to fast CoT responses) and maintaining the capability of instant response on benchmarks. This includes the mathematical benchmarks, such as Spoken-MQA (Wei et al., 2025), and the logical benchmarks, such as SATA-Bench (Xu et al., 2025b). Crucially, in linguistic quality evaluation, our method generates speech with significantly higher fluency than traditional interleaved reasoning models.

2 Related Work

Spoken Language Models. The development of speech interaction technology has witnessed a transition from mechanical responses to human-like natural conversation. In early research, speech interaction systems primarily adhered to a cascaded architecture, a paradigm that strictly divides the interaction process into three independent modules: automatic speech recognition (ASR) (Khedar et al., 2024; Benzeghiba et al., 2007), natural language model for understanding and generation (LLM) (Ouyang et al., 2022; Raffel et al., 2020; Elhady et al., 2025; Kojima et al., 2022a), and text-to-speech (TTS) (Guo et al., 2025). However, the linear transmission between modules leads to significant information loss, particularly as paralinguistic information from speech is almost entirely lost during conversion to text. Moreover, the cumulative latency across processing steps makes it difficult to meet the requirements for real-time interaction in terms of time-to-first-token (TTFT) (Liu et al., 2025). With the advent of the large-scale pre-training era, end-to-end speech models have gradually become a central focus of research. These models attempt to process speech signals within a unified semantic space by discretizing audio into speech tokens or leveraging cross-modal alignment techniques, thereby equipping the model with the ability to directly listen and speak (Wang et al., 2025).

Reasoning Paradigms: From TBS to Interleaved Thinking. To enhance the logical depth of SLMs, researchers introduced the *thinking-before-speaking* (TBS) paradigm, where a model generates a complete hidden CoT before responding. Although accurate, TBS induces prohibitive latency.

To bridge this gap, *thinking-while-speaking* or interleaved reasoning has emerged as a more anthropomorphic paradigm. STITCH (Chiang et al., 2025) utilizes chunked reasoning by exploiting the computational redundancy during speech synthesis. Mini-Omni-Reasoner (Xie et al., 2025) proposes a token-level Thinking-in-Speaking formulation using a hierarchical architecture. More recently, Mind-Paced Speaking (Wu et al., 2025) adopts a dual-brain approach, employing a formulation brain for high-level reasoning to pace a separate articulation brain.

Reinforcement Learning for Reasoning. Reinforcement Learning (RL) has become the dominant paradigm for strengthening LLM reasoning. Prior work demonstrated that models can be guided to interleave thinking and answering through RL, significantly reducing the TTFT (Xie et al., 2026). However, granting models the flexibility often leads to reasoning collapse (Tian et al., 2025), where policies discard explicit reasoning for brevity.

3 Methodology

We introduce the interleaved reasoning framework, a training paradigm designed to bridge the gap between reasoning accuracy and interaction latency. This section outlines our structured data synthesis pipeline, the interleaved supervised fine-tuning strategy, and the subsequent optimization via Group Relative Policy Optimization (GRPO) (DeepSeek-AI, 2024, 2025).

3.1 Interleaved Reasoning

Thinking-while-speaking paradigm enables SLMs to execute reasoning in discrete stages while concurrently delivering verbal summaries. We define a complete response sequence S as follows:

$$S = (T_1, A_1), (T_2, A_2), \ldots, (T_n, A_n), \quad (1)$$

Where T_i represents the i-th thinking segment and A_i represents the corresponding answer segment. In a real-time deployment scenario, these segments serve distinct operational roles:

Background Inference: Each T_i represents a discrete stage of internal logical deduction. These segments provide the rigorous thinking foundation necessary to maintain high-fidelity reasoning, significantly improving response accuracy compared to non-reasoning models.

Foreground Playback: Each A_i serves as a periodical summary that distills the core conclusions of its preceding thinking segment T_i. To ensure a natural and fluent user experience, A_i is specifically optimized to be more concise and colloquial than the raw logic in T_i, mimicking the brevity and tone of human speech. These segments are streamed to a TTS engine for immediate playback.

By interleaved A_i&T_i segments, the generation time for T_{i+1} is masked by the playback time of A_i, the system achieves a seamless interaction flow, effectively concealing the computational latency of complex reasoning while preserving the model's logical rigor.

3.2 Data Construction and Interleaved Refinement

To construct a dataset that supports both high-quality reasoning and fluent speech, we developed a three-stage pipeline to restructure and align raw data into the interleaved reasoning format as shown in Figure 2.

Logical Verification and Oral-Style Summarization Answer Part: We verify the logical integrity of reasoning paths and reconstruct flawed trajectories. These reasoning paths are then distilled into concise, oral-style summaries that preserve essential logic while adopting a natural spoken tone and removing formal writing artifacts. Question Part: Symbolic questions undergo desymbolization into conversational queries to enhance auditory clarity. We then employ CosyVoice2 (Du et al., 2024b) to convert these refined questions into natural-sounding audio inputs.

Semantic Unit Splitting To create natural thinking windows within the voice stream, we split the summary into several small, independent speech units $\{A_1, A_2, \ldots, A_n\}$ based on logical completeness and natural pauses in speech. Each unit is designed to represent a single sub-point or a short sentence that a person would typically say in one breath. This segmentation allows us to insert thinking segments between spoken fragments while maintaining a smooth logical progression.

Controlled-Ratio Thinking Content Construction For each speech unit A_i, we synthesize a corresponding internal thinking segment T_i based on the original logic. To ensure the computational latency is effectively masked during on-device deployment, we enforce a length-ratio constraint. The ratio needs to ensure that under standard on-device inference speeds, the generation of the next reasoning block T_{i+1} is fully covered by the audio playback duration of the current segment A_i. The final

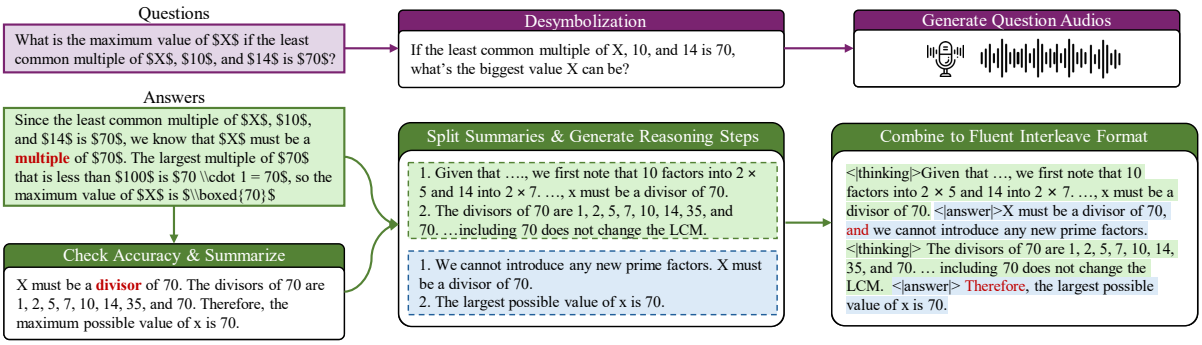

Figure 2: Data Pipeline: (The top line) For questions, the original question first undergoes desymbolization and colloquial normalization to improve its suitability for spoken delivery, followed by speech synthesis via TTS. (The bottom line) For answers, an initial accuracy validation is performed, after which the content is condensed into a summary. This summary is then segmented into semantically coherent units. These segments are subsequently utilized to retrieve corresponding reasoning steps paired with their original answer segments. Finally, the matched thinking-answer pairs are synthesized into an interleaved reasoning response, while simultaneous linguistic refinement ensures the overall coherence and quality of the final output.

output is a structured sequence of (T_i, A_i) pairs that emulate a natural *thinking-while-speaking* cognitive flow.

3.3 Interleaved SFT

Given the structural complexity of frequent logical transitions in the interleaved format, applying RL directly to a base model would result in extremely sparse reward signals and unstable training trajectories. To mitigate this, we implement Supervised Fine-Tuning (SFT) before proceeding to RL optimization. The primary objective of this phase is to align the model's output distribution with the structural requirements of the interleaved reasoning format.

Conventional thinking implementations typically rely on paired delimiters (e.g. `<think>...</think>`) to encapsulate reasoning. In contrast, our framework utilizes singular state-transition flags to minimize sequence overhead. We define two specialized tokens: the trigger `<|thinking|>` signals the model to transition from verbalization to internal logical deduction, and the trigger `<|answer|>` signals a switch from reasoning back to generating a conversational response. This design treats each token as a functional operator that marks the boundary of a state change within the interleaved topology. By using single-token flags instead of paired tags, we effectively compress the overall sequence length and improve inference efficiency—saving valuable tokens in long-form multi-turn dialogues.

The SFT stage serves as a crucial initialization step. By training on the high-quality synthetic data described in the previous section, the model acquires foundational instruction-following capabilities while internalizing the *thinking-answer* alternating pattern. This ensures that when the model enters the RL phase, it already possesses a high degree of format compliance. Establishing this stable baseline prevents the model from collapsing during early RL iterations and significantly enhances the convergence efficiency of the policy gradient optimization.

3.4 Reinforcement Learning

To further optimize the quality of outputs, this paper employs the GRPO algorithm. This algorithm enables policy updates without requiring an extra Critic network by sampling a group of candidate outputs and computing relative rewards.

3.4.1 Reward Function Modeling

The overall reward function R_{total} is designed as a weighted sum of a TA-Balance Reward, an Accuracy Reward, and a Linguistic Quality Reward:

$$R_{total} = \omega_{TA} \cdot R_{TA} + \omega_{Acc} \cdot R_{Acc} + \omega_{LQ} \cdot R_{LQ}, \quad (2)$$

where ω represents the hyperparameter weights for each reward component.

TA-Balance Reward (R_{TA}) The model must strictly adhere to the interleaved reasoning topology. Any violation of the format (e.g., consecutive thinking blocks or missing tags) results in a zero reward. Additionally, to prevent over-thinking on simple tasks and ensure the thinking segments fit within the audio playback window, we introduce a quadratic length penalty. For the i-th thinking segment, its format score r_i is calculated as follows:

$$r_i = \max\left(0, 1 - \left(\frac{L_i - L_{target}}{L_{target}/2}\right)^2\right), \quad (3)$$

where L_i is the word count of the current thinking segment, and L_{target} is the preset word count threshold. The final format reward is the arithmetic mean of all thinking segment scores.

The quadratic decay provides a smoother gradient signal. This soft constraint encourages the model to stay near the L_{target} without being overly penalized for minor deviations, ensuring a stable optimization process.

Accuracy Reward (R_{Acc}) The accuracy reward serves as a hard constraint on logical correctness. we extracts the final prediction y_{pred} from the content following the last <|answer|> token and compares it with the ground truth y_{gt}:

$$R_{Acc} = \mathbb{I}(y_{pred} = y_{gt}) = \begin{cases} 1, & \text{if } y_{pred} = y_{gt} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Linguistic Quality Reward (R_{LQ})

The interleaved reasoning format inherently risks fragmenting the verbal output, potentially leading to semantic disjointedness or lack of narrative flow between separated speech segments. To counteract this, we aim to enforce that the final concatenated response, denoted as:

$$\mathcal{A} = A_1 \oplus A_2 \oplus \cdots \oplus A_n, \quad (5)$$

maintains the coherence of a natural-language sentence. We leverage the linguistic priors of a reference model to guide this process (Chen et al., 2025). The core premise is that a fluent and semantically consistent response should exhibit low perplexity (and thus high likelihood) under the distribution of a frozen reference model. For a given question $\mathcal{Q}$ and a concatenated full answer $\mathcal{A}$, we compute its log-likelihood under the reference distribution p_{ref}:

$$L(\mathcal{A}) = \sum_{t=1}^{|\mathcal{A}|} \log p_{\text{ref}}(a_t \mid \mathcal{Q}, a_{<t}), \quad (6)$$

A higher value of $L(\mathcal{A})$ indicates that the generated sequence aligns closely with natural language patterns learned by the reference model. For a candidate k within a group $\mathcal{G}$, we compute its normalized likelihood score $\tilde{L}^k$ and the Linguistic Quality Reward is formulated as:

$$R_{LQ}^k = \mathbb{I}(k \in \mathcal{C}) \cdot \max\left(0, \beta \cdot \left(\tilde{L}^k - \text{mean}_{\mathcal{G}}(\tilde{L})\right)\right), \quad (7)$$

where $\mathcal{C}$ represents the subset of samples that arrived at the correct final answer, and $\beta > 0$ is a scaling factor. We apply this reward exclusively to correct samples to prevent the model from optimizing for fluency at the expense of logical accuracy. Additionally, a KL penalty is omitted to avoid overly constraining the model's exploratory capability, thereby enabling the integration of this reward at virtually no additional cost relative to the standard GRPO framework. By explicitly rewarding high-likelihood sequences, we encourage the model to bridge the gaps between interleaved segments, ensuring that the final output constitutes a coherent and logically unified narrative suitable for high-quality speech synthesis.

4 Experiments

4.1 Implementation Details

We employ Qwen2.5-Omni-3B (Xu et al., 2025a) as the baseline model to leverage its inherent multimodal capabilities. The fine-tuning process focuses on the thinker module to cultivate interleaved reasoning paths for complex logical and mathematical tasks. The training set comprises approximately 12,000 high-quality samples from K&K (Xie et al., 2024) and MetaMath (Yu et al., 2024a), which are pre-processed into the interleaved reasoning format to ensure a balanced ratio between internal thought segments and audible responses. The training pipeline proceeds in two sequential phases. First, a interleaved SFT phase initializes the interleaved output structure and establishes foundational reasoning habits. This stage runs for 3 epochs with a learning rate of 4×10^{-6} under a cosine decay schedule. It is followed by a RL phase lasting 2 epochs, using a group size of $G = 16$ and a learning rate of 1×10^{-6} .

4.2 Evaluation Metrics

To comprehensively evaluate performance in spoken interaction scenarios, we focus on three core dimensions: Accuracy, Fluency, and Temporal Efficiency.

4.2.1 Accuracy and Generalization

For in-domain evaluation, we use the test sets of K&K and MetaMath. To examine out-of-distribution generalization, we employ Spoken Multi-Question Answering (SMQA) (Wei et al., 2025) and the SATA-Bench (Xu et al., 2025b). SMQA evaluates mathematical reasoning across

four distinct categories, including Short Digit (S) and Long Digit (L) for arithmetic operations, alongside Single-step Reasoning (R_1) and Multi-step Reasoning (R_m) for mathematical logic tasks. Given the varying difficulty and sample sizes, we calculate a weighted aggregate score to represent the overall performance:

$$Score_{total} = \sum_{i \in S, L, R_1, R_m} \frac{N_i}{N_{total}} \cdot Score_i, \quad (8)$$

where N_i represents the number of questions in each branch. We also use SATA-Bench, which covers domains such as law, medicine, and logic reasoning. In this benchmark, the model must correctly identify all valid options in a multiple-choice setting. This benchmark is particularly rigorous because human accuracy is only 17.93%, providing a strong baseline for evaluating advanced reasoning capabilities.

4.2.2 Fluency Assessment

To ensure the interleaved format does not compromise the quality of the audible response, we concatenate the answer segments $\{A_1, A_2, \ldots, A_n\}$ into a single text. We then utilize DeepSeek-V3 as an automated judge to score the overall fluency on a scale of 0 to 2:

0 (Incoherent): Logical contradictions exist between segments. For example, a model might first state "The total loss is 48 units" and later claim "Therefore, the net loss is 32 units" without reconciliation.

1 (Fluent but Disjointed): Logical flow is correct, but linguistic transitions between segments are abrupt. For example, the reply "She spends $10 on orange creamsicles. She spends $4.50 on ice cream sandwiches. She spends $14.50 in total." might have a correct solving process, but there is no smooth transition between sentence.

2 (Excellent): Both logic and expression are seamless and coherent.

4.2.3 Temporal Efficiency

Immediate responsiveness is critical for the user experience in real-time conversational AI. In our evaluation, we employ Instant Response as a categorical metric to denote whether a model facilitates immediate streaming interaction or necessitates high-latency pre-computation.

4.3 Main Result

We conducted a comprehensive evaluation of the proposed *interleaved reasoning* paradigm and various baseline models. The performance across logical and mathematical reasoning benchmarks is detailed in Table 1.

4.3.1 Models and Baselines

To provide a robust comparison, we categorize the compared methods into four distinct paradigms based on their reasoning and response patterns:

Thinking: This paradigm follows the traditional *thinking-before-speaking* approach. **Zero-shot** utilizes the vanilla Qwen2.5-Omni-3B, prompted to generate a full reasoning chain before producing the answer. **SFT+RL** is further fine-tuned and optimized with reinforcement learning to improve the performance.

Direct: This paradigm constrains the model to output only the final answer without any explicit reasoning steps.

Fast CoT: This paradigm aims for providing a concise, *fast-thinking* summary. The **Zero-shot** variant is elicited through specific prompting, whereas **SFT+RL** is specifically trained to optimize this fast-thinking behavior.

Interleaved: This paradigm integrates reasoning and speech into a single, unified stream. **Mini_OR** (Xie et al., 2025) is an external state-of-the-art baseline for interleaved reasoning; **SFT** is trained with supervised fine-tuning to master the interleaved structure; **InterRS** is our final model, optimized from the SFT base using our proposed method.

4.3.2 Analysis of Key Findings

The result suggests that our proposed InterRS model effectively reconciles reasoning rigor with real-time interaction. It achieves a highly competitive average score of 49.33, closely approaching the accuracy of monolithic Thinking SFT+RL model of 51.10. Crucially, InterRS attains this while maintaining instant response capability. In contrast, the Thinking baseline, despite its marginal accuracy lead, requires full pre-computation, rendering it unsuitable for smooth conversation. This demonstrates that our InterRS preserves over **96.5%** of the analytical capability of Thinking SFT+RL model without incurring its latency costs.

Furthermore, the comparison highlights a trade-off: models of Thinking, Direct and Fast CoT that prioritizing speed suffer significant accuracy drops on complex tasks. Conversely, the Thinking models that achieve high accuracy sacrifice all real-time responsiveness. InterRS uniquely demonstrates

Type	Method	Logic		Mathematical		Avg. ↑	Inst. Resp.
		SATA-Bench	KK-audio	SMQA	Meta-audio		
Thinking	Zero-shot	0.30	5.57	70.11	52.60	32.15	×
	SFT+RL	24.00	42.14	**76.74**	**61.50**	**51.10**	×
Direct	Zero-shot	0.20	6.57	42.15	10.50	14.86	√
Fast CoT	Zero-shot	0.10	6.57	69.40	19.70	23.94	√
	SFT+RL	**25.20**	28.57	53.19	40.70	36.92	√
Interleaved	Mini_OR	-	-	69.01	-	-	√
	SFT	17.50	39.14	73.01	59.10	47.19	√
	InterRS	21.10	**42.57**	74.43	59.20	49.33	√

Table 1: **Main Result on Logical and Mathematical Benchmarks.** We report accuracy scores (↑) across four datasets. The **bolded result** indicates the highest score, and the underlined result indicates the second highest score. Results demonstrate that our InterRS achieves competitive accuracy while maintaining the capability of instant response, bridging the gap between reasoning and real-time interaction.

how high reasoning accuracy and low latency can be jointly optimized for conversational AI.

5 Ablation Studies

In this part we train and test the models only on mathematical datasets.

5.1 Analysis of Length and Interleaving Controls

We evaluated the impact of our TA-Balanced rewards against three alternative configurations to understand how different optimization objectives influence reasoning behavior. The **Thinking** model follow the *thinking-before-speaking* format. The **Inter** model enforces correct tag structures but permits unconstrained thinking segment lengths. The **Inter. w. Seg** rewards the model for reducing the total number of interleaved transitions under the correct interleaved format, which may alleviate over thinking on simple questions. Our **Inter. w. TA** encourages the model to generate thinking segment under target length and keep a balanced length ratio between thinking segments and answer segments. We examine the stability of individual thinking segments. Figure 4 reveals that the Thinking model is unsuitable for real-time interaction due to excessive latency, evidenced by an IQR of 42 tokens. The Inter. w. Seg and Inter baseline models lack rhythmic consistency, as they consolidate reasoning into obstructive blocks with IQRs of 26 and 12 tokens, respectively. **The Inter. w. TA model achieves the most refined control**, yielding the narrowest IQR of only 11 tokens and the lowest median. This statistical stability effectively eliminates the risk of long-duration reasoning blocks, providing the predictable and fluent interaction cadence required for a controlled real-time framework.

Model	Fmt	Seg	TA	Acc	Len
Thinking	√	×	×	**69.40**	120.87
Inter	√	×	×	68.63	128.64
Inter w.Seg	√	√	×	69.02	123.61
Inter w.TA	√	×	√	68.65	**104.88**

Table 2: Ablation studies across total thinking segment lengths and interleaving times controls.

No.	Smooth	LQ Rew	Acc	Fluency
(1)	×	×	68.65	1.65
(2)	√	×	68.32	1.74
(3)	√	√	**69.29**	**1.83**

Table 3: Ablation study of **conversational smoothing and Linguistic Quality reward**.

According to Table 2, while the Inter. w. Seg configuration yields a slightly higher accuracy, it results in a significantly higher total thinking token count. In contrast, our Inter. w. TA model maintains competitive performance while achieving the shortest total thinking length.

This reduction in total thinking length significantly enhances the system's inference efficiency. A lower total token count translates to higher reasoning density, where the model learns to identify the most direct logical path while minimizing redundant computation.

5.2 Impact of Data Refinement and LQ Reward

We further investigate the contributions of our data refinement pipeline and the Linguistic Quality Reward, with all configurations built upon the TA-Balanced Reward foundation. Configuration (1) serves as the baseline, corresponding to the Inter.

Question:
If Djibo is currently 17 years old and the sum of his age and his sister's age was 35 five years ago, what is the current age of Djibo's sister?

Config (1): Unrefined data

T1: Five years ago, Djibo's age was 17 − 5 = 12 years.
T2: Let Djibo's sister's age five years ago be x. Their combined age then was 12 + x.
T3: We know their combined age five years ago was 35, so 12 + x = 35.
T4: Solving 12 + x = 35 gives x = 23. This is her age five years ago.
T5: Her current age is 23 + 5 = 28 years old.

A1: Five years ago, Djibo was 12 years old.
A2: Five years ago, their combined age was 12 + x.
A3: The equation is 12 + x = 35.
A4: Her age five years ago was 23.
A5: She is currently 28 years old.

Config (3): Refined data and LQ Reward

T1: Five years ago, Djibo was 17 - 5 = 12 years old. His sister's age at that time was unknown, but we know the sum of their ages back then was 35.
T2: If Djibo was 12 and they were 35 years old in total, then his sister's age five years ago was 35 - 12 = 23 years old.
T3: Since five years have passed since then, we add 5 years to her age from five years ago to find her current age: 23 + 5 = 28.

A1: Five years ago, Djibo was 12 years old, and together they were 35 years old.
A2: At that time, his sister was 23 years old.
A3: Currently, Djibo's sister is 28 years old.

Figure 3: Comparison of responses from Configuration (1) and Configuration (3) for a multi-step reasoning task.

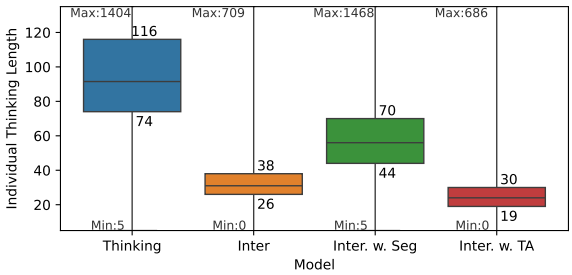

Figure 4: Distribution of **individual thinking segment lengths** across different model configurations.

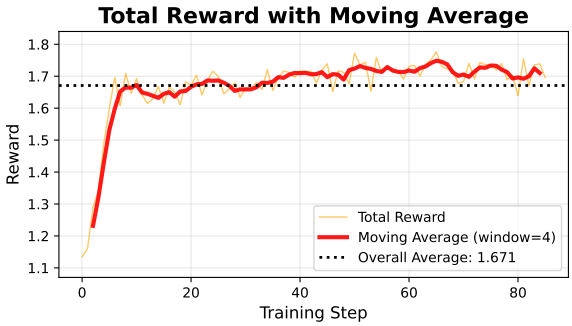

Figure 5: Rewards of our InterRS in training process.

w. TA model from the preceding analysis.

According to Table 3, the transition from Configuration (1) to (2) demonstrates that training on our refined data significantly enhances the naturalness of the spoken output, increasing the fluency score from 1.65 to 1.74. The qualitative shift toward more human-like verbal expressions confirms the effectiveness of the data refinement pipeline. Building upon the smoothed data, **InterRS** (Configuration 3) introduces the Linguistic Quality Reward, which **leads to the best overall performance**, achieving a peak fluency score of 1.83 and an improved accuracy of 69.29. This progression confirms that explicitly rewarding the semantic alignment between internal reasoning and audible segments is essential. This ensures that the fragmented response constitutes a coherent, logically unified answer rather than a series of disjointed statements. The reward curve is shown on Figure 5.

Figure 3 illustrates the qualitative advantages of InterRS through a case study. Configuration (1) produces fragmented, isolated segments that lack natural flow and introduces an undefined variable x directly into the audible output. Conversely, Configuration (3) achieves superior semantic continuity by synthesizing reasoning steps into cohesive, fluent sentences. By employing transitional phrases (e.g., "At that time") to anchor logical inferences, it ensures mathematical accuracy while optimizing the delivery for human-like speech communication.

6 Conclusion

In this work, we present a novel training method designed to achieve instant response while maintaining high-quality thinking in spoken language models. By using the delay in speech generation to mask thinking tokens, our approach eliminates the prohibitive latency typical of conventional *thinking-before-speaking* models.

Our approach preserves high analytical accuracy while ensuring instantaneous response capability. We achieve fine-grained control over thinking lengths through the TA-Balanced reward. By eliminating the risk of unexpectedly long reasoning blocks, this mechanism ensures a stable interaction cadence. Furthermore, the synergy between conversational data refinement and the Linguistic Quality reward allows the model to achieve SOTA fluency and accuracy. These optimizations ensure that the interleaved response stream constitutes a coherent, logically unified answer rather than a series of disjointed statements, validating the feasibility of integrating high-level reasoning with temporally efficient speech generation.

7 Limitations

While our model achieves instant response times resembling a spoken-language instruct model, the current framework evaluates the interaction in a primarily static turn-taking format. Real-world human communication is highly dynamic, often involving interruptions or abrupt topic shifts. How the interleaved reasoning segments adapt or truncate when the external speech flow is interrupted by the user is an area that requires dedicated architectural extensions in future work.

Another open question concerns the model's ability to handle queries that require little or no deliberate reasoning. While our analysis shows that the model naturally reduces thinking length for simple arithmetic tasks, a systematic evaluation of providing instantaneous, thought-free responses to trivial questions is left for future work. Integrating fast-thinking and slow-thinking capabilities into a unified, adaptive architecture is a promising direction for further improving both efficiency and user experience.

References

Mohamed Benzeghiba, Renato de Mori, Olivier Deroo, Stéphane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, Richard Rose, Vivek Tyagi, and Christian Wellekens. 2007. Automatic speech recognition and speech variability: A review. *Speech Commun.*, 49(10-11):763–786.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Senter, Natalia Elena, Neil Zeghidour, Marco Tagliasacchi, Matt Sharifi, Neil Christian, and Dominik Roblek. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2534.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. 2025. GRPO-CARE: consistency-aware reinforcement learning for multimodal reasoning. *CoRR*, abs/2506.16141.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. 2025. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models.

Yunfei Chu and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037.

Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. 2024a. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. *CoRR*, abs/2409.19510.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent abilities of large language models under continued pre-training for language adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 32174–32186. Association for Computational Linguistics.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2025. Recent advances in discrete speech tokens: A review. *CoRR*, abs/2502.06490.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Inf. Fusion*, 109:102422.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022a. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Takeshi Kojima and 1 others. 2022b. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*.

Kushal Lakhotia and 1 others. 2021. Generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*.

Zhanxun Liu, Yifan Duan, Mengmeng Wang, Pengchao Feng, Haotian Zhang, Xiaoyu Xing, Yijia Shan, Haina Zhu, Yuhang Dai, Chaochao Lu, Xipeng Qiu, Lei Xie, Lan Wang, Nan Yan, Zilong Zheng, Ziyang Ma, Kai Yu, and Xie Chen. 2025. X-talk: On the underestimated potential of modular speech-to-speech dialogue system.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *CoRR*, abs/2501.07246.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Paul K Rubenstein and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Harvey Sacks and 1 others. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-audio-r1 technical report.

Qing Wang, Zehan Li, Yaodong Song, Hongjie Chen, Jian Kang, Jie Lian, Jie Li, Yongxiang Li, and Xuelong Li. 2026. A unified spoken language model with injected emotional-attribution thinking for human-like interaction. *arXiv preprint arXiv:2601.04960*.

Siyin Wang, Zengrui Jin, Changli Tang, Qiujia Li, Bo Li, Chen Chen, Yuchen Hu, Wenyi Yu, Yixuan Li, Jimin Zhuang, Yudong Yang, Mingqiu Wang, Michael Han, Yifan Ding, Junwen Bai, Tom Ouyang, Shuo yiin Chang, Xianzhao Chen, Xiaohai Tian, and 10 others. 2025. Towards general auditory intelligence: Large multimodal models for machine listening and speaking.

Chengwei Wei, Bin Wang, Jung-Jae Kim, and Nancy F. Chen. 2025. Towards spoken mathematical reasoning: Benchmarking speech-based models over multi-faceted math problems. *CoRR*, abs/2505.15000.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Maeda, Ed Chi, Sharan Narang, Aakriti Chowdhery, and Quoc V Le. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Donghang Wu, Haoyang Zhang, Jun Chen, Xiangyu, Zhang, Hexin Liu, Eng Siong Chng, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Mind-paced speaking: A dual-brain approach to real-time reasoning in spoken language models.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *CoRR*, abs/2410.23123.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. 2026. Interleaved reasoning for large language models via reinforcement learning.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. 2025. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models.

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

Weijie Xu, Shixian Cui, Xi Fang, Chi Xue, Stephanie Eckman, and Chandan K. Reddy. 2025b. SATA-BENCH: select all that apply benchmark for multiple choice questions. *CoRR*, abs/2506.00643.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024a. Metamath: Bootstrap your own mathematical questions for large language models.

Tianshu Yu and 1 others. 2024b. Towards effective reasoning in spoken dialogue systems. *Proceedings of ACL*.

Dong Zhang, Shimin Li, Nan Zhang, Xuan Zhan Chen, Zuxuan Wang, Bingzheng Zheng, Shumin Kang, Jiahui Lin, Jiaheng Zhang, Chenhai Feng, and 1 others. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

A Determination of Target Thinking Length

The target thinking segment length L_{target} depends on the generation speed V_{gen}, the voice playback duration $T_{\text{play},i}$, and the TTS latency L_{tts}. To avoid audible pauses, the thinking token count must satisfy:

$$N_{\text{think},i+1} \times V_{\text{gen}} + L_{\text{tts}} \leq T_{\text{play},i}. \quad (9)$$

In practice, L_{target} should be chosen per deployment scenario according to this constraint. Our experimental setup uses the following values:

- $V_{\text{gen}} = 30$ ms/token,
- $L_{\text{tts}} = 300$ ms,
- $V_{\text{play}} = 400$ ms/word,
- $k = 1.3$ (token-to-word conversion ratio).

To ensure robustness for short responses (e.g., five words), we compute the maximum allowed thinking tokens and the corresponding target length:

$$N_{\text{think}} \leq \frac{5 \times 400 - 300}{30} \approx 56.6 \text{ tokens}, \quad (10)$$

$$L_{\text{target}} = N_{\text{think}}/k \approx 43.6 \text{ words}. \quad (11)$$

We tested three L_{target} settings (30, 40, 50) and report accuracy and timeout metrics in Table 4. The results demonstrate that the model's accuracy is largely insensitive to the choice of L_{target} within a reasonable range. The performance variation across tasks and difficulty levels remains minimal, indicating that our method does not rely on task-specific hyperparameter tuning and exhibits strong general applicability across diverse reasoning scenarios.

Table 4: Effect of L_{target} on accuracy and latency. TO Count is the abbreviation for timeout count, and TO avg. is the abbreviation for average timeout duration.

L_{target}	Acc. (%)	TO Count	TO Avg. (ms)
30	67.13	3	210.0
40	69.29	10	267.6
50	68.78	12	294.0

B Latency Evaluation Results

In modern SLM frameworks, text token generation is typically much faster than audio frame synthesis and playback. This speed mismatch creates idle periods during which the reasoning engine waits for the audio buffer to clear — a waste of computational resources. Our interleaved reasoning paradigm repurposes these idle periods as opportunities for thinking. By precisely aligning reasoning steps with audio rendering lags, the method transforms otherwise wasted waiting time into productive computation. As a result, the peak computational demand per second never exceeds the hardware's baseline capacity for standard SLM tasks, avoiding audio stuttering without requiring additional resources.

We measured the average end-to-end latency for each model using a faster generation speed of $V_{\text{gen}} = 15.8$ ms/token (Table 5). Our InterRS model reduces latency from 1020 ms to 60 ms compared to the think-before-speak (Thinking) model, while retaining 96.5% of its accuracy.

Table 5: Accuracy and average latency across models.

Model	**Acc. (%)**	**Latency (ms)**
Fast CoT	36.92	0
Thinking	51.10	1020
Interleaved-SFT	47.19	267.8
InterRS (Ours)	49.33	60