Dual-Reasoner: Bridging Interleaved Atomicity and Streaming Latency via Thinking-while-Talking

Yangzhuo Li [♣ ♢ *] **Shengpeng Ji** [♠ *] **Yifu Chen** [♠ *] **Tianle Liang** [♠]
Haoyu Yang **Junbo Li** **Jun Fang** **Lin Li** [♣] **Qingyang Hong** [♣ †]
[♣]Xiamen University [♢]Jilin University [♠]Zhejiang University
liyangzhuo49@gmail.com, qyhong@xmu.edu.cn

Abstract

Integrating explicit Chain-of-Thought (CoT) into end-to-end spoken dialogue models enhances intelligence but incurs prohibitive latency. While the **"Thinking-while-Talking"** paradigm alleviates this delay, it fundamentally compromises block atomicity, severing the logical connection between interleaved thought and speech. To address this, we present **Dual-Reasoner**, employing a **Streaming Masking Mechanism** underpinned by our **Dual-Think-30k** dataset to guarantee uninterrupted audio streaming. Crucially, to strictly align the fragmented thinking blocks to service speech generation, we introduce the **Atomic-Consistency Restoration** framework. To secure comprehensive capabilities in high-difficulty reasoning, this mechanism utilizes a quadruple-constraint system to reconstruct logical atomicity, ensuring that "think" chunks act as a rigorous anchor for "talk" outputs. Experimental results demonstrate that Dual-Reasoner achieves comprehensive reasoning enhancements within ultra-low latency constraints: it elevates the VoiceBench score from 67.24 to 73.41 over the baseline, while significantly reducing the Time-to-First-Audio (TTFA) from 20.35s to 3.65s and the Real-Time Factor (RTF) from 7.04 to 1.05.

1 Introduction

Large Language Models (LLMs) elicit multi-step reasoning via Chain-of-Thought (CoT) prompting (Kojima et al., 2022; Wang et al., 2023; Wei et al., 2022). By explicitly generating intermediate reasoning steps, CoT significantly enhances logical coherence and interpretability, demonstrating improved performance in complex reasoning tasks. Given that Spoken Dialogue Systems (Ji et al., 2024; Xu et al., 2025; KimiTeam et al., 2025; Team et al., 2025) offer natural Human-Computer Interaction, current research focuses on integrating CoT reasoning to bolster capabilities in both complex reasoning tasks and robustness challenges specific to speech interaction (Xie et al., 2025; Ma et al., 2025a; Goel et al., 2025; Wang et al., 2025a). However, applying CoT to Spoken Dialogue Models faces a critical challenge: the Real-Time Interaction Bottleneck. The conventional "Thinking-then-Talking" (Wei et al., 2022) paradigm, requiring full reasoning prior to synthesis, incurs excessive Time-to-First-Token (TTFT) and unnatural silence, thereby violating the requirements for real-time interaction.

Current optimizations for End-to-End Spoken CoT fall into two categories. The first focuses on compressing thought length or accelerating generation (Ning et al., 2024; Hsieh et al., 2023), yet fails to overcome the **"Thinking-then-Talking"** bottleneck under ultra-low latency constraints. The second, exemplified by **"Thinking-while-Talking"** (Chiang et al., 2025b,a), significantly reduces TTFT by interleaving "think" and "talk" chunks. However, these approaches typically rely on rigid mechanical truncation based on fixed token counts, resulting in intra-chunk semantic incompleteness and logical disjointedness. Consequently, existing "Thinking-while-Talking" paradigms face three critical challenges: 1) The Sequence Length Allocation Dilemma: While sub-optimal absolute lengths necessitate a trade-off between TTFT and model performance, relative imbalance poses a critical risk: if subsequent thought generation exceeds preceding speech playback, it triggers client-side Audio Buffer Underflow. 2) Modality Switching Cost: Frequent state switching disrupts attentional continuity, inducing context fragmentation and memory degradation. This hinders maintaining a global reasoning trajectory, rendering multi-step CoT prone to irreversible logical discontinuity and severe hallucinations. 3) Semantic Integrity and Granularity Mismatch: Mechanical truncation compromises the atomicity of CoT steps, yielding semantically

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 4081–4105
July 2-7, 2026 ©2026 Association for Computational Linguistics

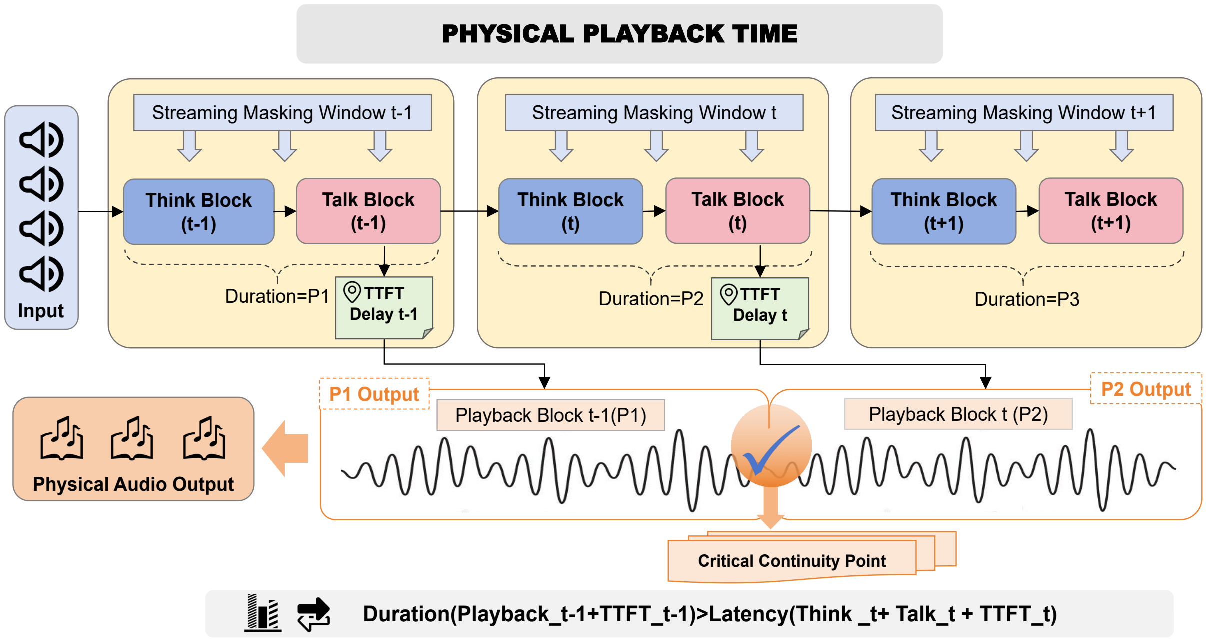

Figure 1: Schematic illustration of the **Streaming Masking Mechanism**. Operating within the "Thinking-while-Talking" paradigm, the model sequentially generates Think and Talk Blocks. To maintain seamless Physical Audio Output, the duration of the preceding Playback Block t-1 must effectively "mask" the latency of the **Streaming Masking Window t**. This temporal overlap guarantees that the subsequent audio segment P_2 is prepared before the current playback P_1 concludes at the Critical Continuity Point.

incomplete fragments rather than functional closed loops. This precludes treating chunks as discrete reasoning steps, failing to robustly support complex problem solving via explicit Sub-task Decomposition.

To enable logically coherent "Thinking-while-Talking," we propose the Dual-Reasoner architecture for Streaming S2S tasks alongside the Dual-Think-30k dataset. As illustrated in Figure 1, Dual-Reasoner employs a **Streaming Masking Mechanism** to solve physical latency. This mechanism enforces a constraint: the sum of decoding and playback duration of the current talking chunk must strictly cover the cumulative latency of the subsequent reasoning step, speech generation, and first-frame audio synthesis. By leveraging the playback window to "mask" computation latency, the model ensures continuous audio output. As shown in Figure 2, to address the logical discontinuity inherent to chunked generation, we introduce the **Atomic-Consistency Restoration (ACR)** framework. Moving beyond simple reward aggregation, this system utilizes four synergistic constraints to reconstruct logical atomicity: it secures intra-block integrity via Semantic Completeness, strictly aligns speech with preceding thoughts through Internal-External Consistency, bridges distal reasoning gaps with Logic Flow, and guarantees the rigorous accuracy and high-quality expression of the final output via Multimodal Outcome. This framework ensures that fragmented "think" blocks act as a rigorous cognitive foundation specifically servicing "talk" outputs. We also constructed the Dual-Think-30k dataset to facilitate dynamic switching between "Fast and Slow" systems. Integrating complex logic samples—constructed via the R-P-T-S (Role-Problem-Task-Symbolic) framework—with casual conversational data, it supports the architectural demands. Consequently, Dual-Reasoner transcends the "Thinking-then-Talking" paradigm in terms of latency. By leveraging optimized think-talk synergy, it not only achieves comprehensive reasoning enhancements but also drastically reduces the Real-Time Factor (RTF), successfully reconciling the conflict between deep cognition and instant delivery. In summary, our main contributions are as follows:

- **Dual-Reasoner**: We propose a novel architecture tailored to the "Thinking-while-Talking" paradigm, capable of generating responses characterized by both sophisticated reasoning capabilities and logical rigor within strict latency constraints.

- **Atomic-Consistency Restoration Framework**: We introduce a novel optimization framework designed to reconstruct Think-Talk Atomicity within the "Thinking-while-Talking" paradigm. Moving beyond simple reward aggregation, this mechanism utilizes four synergistic constraints to align fragmented reasoning with speech generation, ensuring that the interleaved stream maintains the logical integrity and depth of a unified cognitive process.

- **Dual-Think-30k Dataset**: We release a speech Chain-of-Thought (CoT) annotated reasoning dataset that integrates the "Fast and Slow" thinking paradigm. This dataset effectively enhances the overall performance of speech models in both logical reasoning and general interaction tasks.

2 Related Work

Research on Spoken Chain-of-Thought (CoT) primarily focuses on transferring LLM reasoning capabilities to audio modalities via the "Thinking-then-Talking" paradigm. Audio-CoT (Ma et al., 2025a) utilizes zero-shot or few-shot prompting to elicit reasoning in existing models. To enhance robustness, methods like Audio-Reasoner (Xie et al., 2025), Mellow (Deshmukh et al., 2025), and Audio Flamingo 3 (Goel et al., 2025) employ Supervised Fine-Tuning (SFT) to distill multi-step reasoning from high-quality synthetic datasets. Recently, Reinforcement Learning was introduced in models like Step-Audio-R1 (Tian et al., 2025) and PAPO (Wang et al., 2025b) to optimize reasoning paths explicitly. Despite their effectiveness in complex tasks, these sequential models inherently suffer from the "Real-Time Interaction Bottleneck," as they must complete full reasoning before initiating speech synthesis, resulting in prohibitive TTFT.

To address latency, recent works have explored interleaved generation. Shanks (Chiang et al., 2025a) introduces a "Listening-while-Thinking" approach, while **STITCH** (Chiang et al., 2025b) proposes a "Thinking-while-Talking" framework that reduces latency by alternating between thinking and talking tokens. However, these methods rely on mechanical truncation based on fixed token counts or heuristics. This rigid segmentation leads to intra-chunk semantic incompleteness and logical disjointedness. Unlike these approaches, our work focuses on restoring the atomicity of reasoning steps within a streaming architecture, achieving both semantic integrity and low-latency response via a constraint-aware ACR Framework.

3 Method

3.1 Overview

We formally formulate the realization of "Thinking-while-Talking" not merely as a latency reduction strategy, but as a structural reconstruction problem aiming to **reconstruct Logical Atomicity** within the constraints of **Ultra-low Latency Interaction**. While the "Thinking-while-Talking" paradigm inherently mitigates response delay, existing models confront a fundamental "Spatiotemporal Mismatch": physically, the client-side Audio Buffer demands a continuous data stream to prevent underflow; cognitively, the fragmented generation of Chain-of-Thought (CoT) severs the logical connection between reasoning and expression. To resolve this fragmentation without compromising the immediate response capability, our methodology is constructed upon two orthogonal pillars:

Physical Continuity via Dual-Reasoner: To guarantee uninterrupted playback under ultra-low latency constraints, we propose the Dual-Reasoner architecture centered on a Streaming Masking Mechanism. This transforms the latency bottleneck into a manageable physical inequality: by enforcing a constraint where the playback duration of preceding talking segments masks the computational overhead of subsequent reasoning blocks, we ensure continuous audio output at the physical layer.

Logical Atomicity via Atomic-Consistency Restoration: To secure comprehensive reasoning capabilities despite compromised block integrity, we introduce the Atomic-Consistency Restoration (ACR) framework at the cognitive layer. Utilizing a quadruple-constraint system, this mechanism structurally reconstructs Think-Talk Atomicity by enforcing a strict causal dependency between interleaved blocks. It unifies the fragmented generation process, mandating that "think" chunks function as rigorous logical anchors for "talk" outputs, thereby effectively mitigating the hallucinations and memory degradation induced by frequent mode switching.

The following subsections detail the implementation of these mechanisms and the data foundation supporting this paradigm.

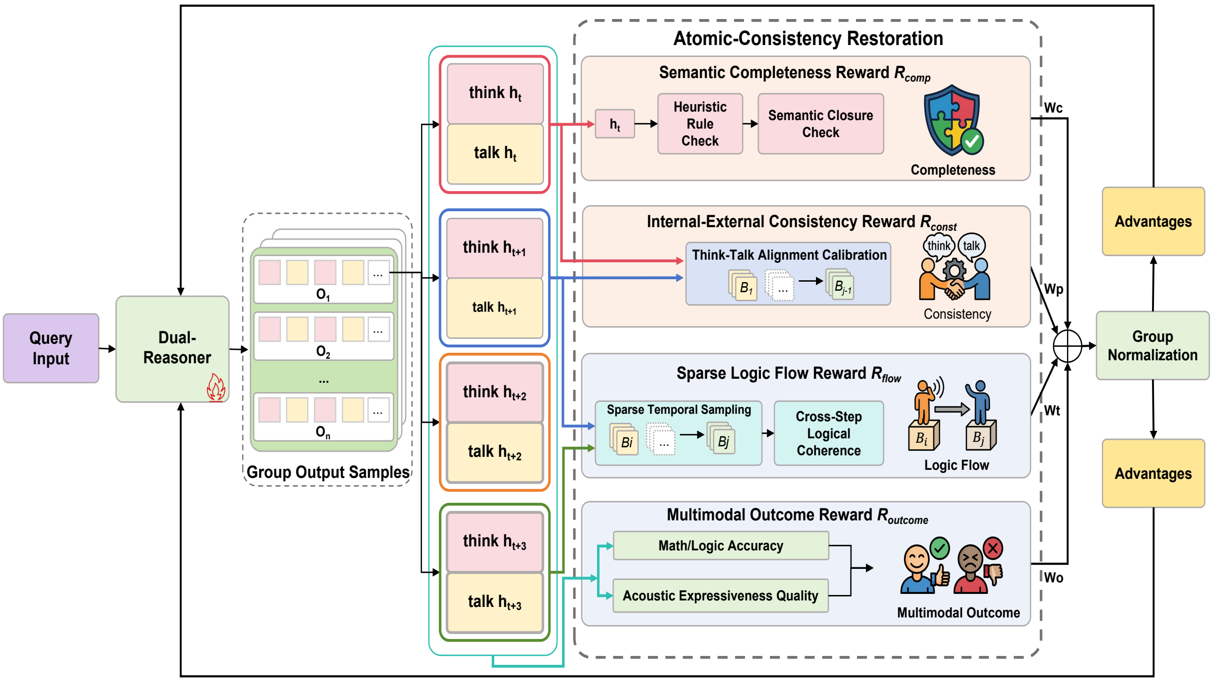

Figure 2: Schematic overview of the **Atomic-Consistency Restoration (ACR)** framework. The policy model generates interleaved think-talk trajectories, structurally optimized via four synergistic constraints: Semantic Completenes establishes intra-block atomicity, Internal-External Consistency enforces strict causal dependency, Sparse Logic Flow guarantees global logical continuity and Multimodal Outcome provides final textual-acoustic dual-validation. These signals are aggregated via Group Normalization for gradient-based optimization.

3.2 Dual-Reasoner

3.2.1 Generative Paradigm

Dual-Reasoner adapts the **Step-Audio-2-mini-Think** (Wu et al., 2025) architecture into a streaming paradigm, producing an interleaved sequence $O = \{Think_i, Talk_i\}_{i=1}^{N}$. To operationalize this, we enforce strict formatting where thinking is encapsulated in <think> tags, and talk chunks utilize <tts-pad> with a fixed 1:4 text-to-acoustic ratio. First, regarding the relative length constraints, we introduce a **Streaming Masking Mechanism** to ensure continuous playback. This mechanism leverages the preceding block's playback duration to mask the current block's computational latency:

$$\begin{aligned}\mathcal{D}(\mathrm{P}_{t-1}) + \mathcal{L}(\mathrm{TTFT}_{t-1}) \\ > \mathcal{L}(\mathrm{Think}_t + \mathrm{Talk}_t + \mathrm{TTFT}_t)\end{aligned} \tag{1}$$

Here, P denotes Playback, while $\mathcal{D}(\cdot)$ and $\mathcal{L}(\cdot)$ represent duration and latency operators. The inequality balances the available time budget (LHS)—comprising the previous block's playback and decoding—against the cumulative startup latency (RHS) required for the current block's reasoning and acoustic initiation. By substituting generation rates based on empirical data in Appendix C, we simplify these physical boundaries into a linear relative token relationship: the thinking length must be strictly controlled within 0.5 times the talking length ($N_{\text{think}} < 0.5 \cdot N_{\text{talk}}$).

Second, regarding the absolute sequence length, we perform a discretized optimization within the feasible region established above. Guided by the statistical prior of mean sentence length $\mathbb{E}[|S|] \in [20, 30]$, we evaluated a gradient configuration set $\mathcal{C}$. Experimental results demonstrate that the configuration $\mathcal{C}_{opt} = (60, 125)$ yields robust performance across evaluated metrics (see Appendix E). This setting semantically maps to a capacity interval of $\mathcal{K}_{think} \in [3\mathbb{E}[|S|], 5\mathbb{E}[|S|]]$, confirming that a window of 3-5 complete sentences effectively mitigates semantic fragmentation while converging to an balanced trade-off between reasoning depth and interaction latency.

3.2.2 Atomic-Consistency Restoration Framework

To structurally reconstruct the **Logical Atomicity** of the interleaved stream and prevent cognitive degradation, we implement the **Atomic-Consistency Restoration (ACR)** framework within the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024). Moving beyond simple reward aggregation, ACR is designed to enforce a unified cognitive process over physically fragmented chunks. As illustrated in Figure 2, we leverage GPT-4o[1] and GPT-4o-audio[2] to evaluate the streaming trajectory $\mathcal{C} = \{Block_1, Block_2, ..., Block_T\}$. The synergistic objective is formalized as a weighted integration of four structural constraints:

$$\begin{aligned} R_{\text{total}} = w_{\text{const}} \cdot R_{\text{const}} + w_{\text{comp}} \cdot R_{\text{comp}} \\ + w_{\text{flow}} \cdot R_{\text{flow}} + w_{\text{out}} \cdot R_{\text{out}} \end{aligned} \tag{2}$$

where $\mathbf{w} = [w_{\text{const}}, w_{\text{comp}}, w_{\text{flow}}, w_{\text{out}}]^\top$ denotes the vector of hyperparameter weights coordinating the restoration process.

1) Semantic Completeness Reward (R_{comp}): Serving as the foundational unit of restoration, this component establishes Intra-Block Atomicity by enforcing the structural integrity of individual blocks. By applying a two-stage verification—Heuristic Rule Checks ($\mathcal{P} = \{., ?, !\}$) followed by Semantic Closure Checks (see Appendix D.1)—we ensure that every physical chunk functions as a cognitively autonomous unit. This prevents the generation of "open loops" or truncated fragments (e.g., "Therefore the result is"), ensuring that each block is semantically self-sufficient before integration.

2) Internal-External Consistency Reward (R_{const}): To reconstruct Think-Talk Atomicity, this constraint enforces a strict causal dependency between interleaved blocks. Governed by the "Logical Alignment Auditor" (see Appendix D.2), it verifies that the current speech (h_t^{talk}) is a rigorous projection of the thought history ($h_{0:t}^{\text{think}}$). This mechanism penalizes "Hallucinated Increments" where speech outpaces reasoning (e.g., articulating "Step A" before deducing "Step A"), mandating that the "Talk" block must be logically anchored in the preceding "Think" block.

3) Sparse Logic Flow Reward (R_{flow}): To bridge the cognitive gaps created by streaming interleaving, this component guarantees Global Logical Continuity beyond local context. Utilizing Sparse Temporal Sampling (see Appendix D.3), it audits non-adjacent block pairs (B_i, B_j) $(i \ll j)$ to detect long-range contradictions. By penalizing distal variable conflicts (e.g., B_j assuming $x = 10$ after B_i defined $x = 5$), it ensures the fragmented stream maintains the rigorous adherence to antecedent premises characteristic of a unified reasoning chain.

[1] https://platform.openai.com/docs/models/gpt-4o

[2] https://platform.openai.com/docs/deprecations#2025-06-10-gpt-4o-audio-preview-2024-10-01

4) Multimodal Outcome Reward (R_{out}): Acting as the final quality gate, this component provides Final Outcome Dual-Validation for the reconstructed trajectory. Specifically, it directly rewards the accuracy of the complete textual answer and the naturalness of the generated audio (see Appendix D.4). This enforces a high-standard synergy: it rewards correct solutions synthesized with natural, context-aware prosody, while significantly penalizing outputs that are factually incorrect or lacking prosodic variation.

3.2.3 Algorithmic Workflow

To optimize streaming strategies, we employ GRPO. Given a query x, the policy π_θ samples a group of G independent trajectories $\{C^{(k)}\}_{k=1}^G$, where each $C^{(k)}$ is composed of sequential atomic blocks $B_t = (h_t^{\text{think}}, h_t^{\text{talk}})$. The total reward $R_{\text{total}}^{(k)}$ is computed as the weighted aggregation of the four aforementioned reward dimensions:

$$R_{\text{total}}^{(k)} = \sum_{d \in \mathcal{M}} w_d \cdot R_d(C^{(k)}) \tag{3}$$

We employ group-level normalization to compute the Advantage $\hat{A}^{(k)}$ for the k-th trajectory:

$$\hat{A}^{(k)} = \frac{R_{\text{total}}^{(k)} - \mu(R_{\text{total}})}{\sigma(R_{\text{total}}) + \epsilon} \tag{4}$$

where μ and σ denote the mean and standard deviation of returns within group $\mathcal{G}$. We update parameters θ by maximizing the Surrogate Objective, which integrates PPO clipping and a KL divergence penalty relative to the reference policy π_{ref}:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, C \sim \pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G} \sum_{k=1}^{G} \Big(\mathcal{L}_{\text{clip}}(\rho_k, \hat{A}^{(k)}) \\ - \beta D_{\text{KL}}(\pi_\theta(C^{(k)}|x) \parallel \pi_{\text{ref}}(C^{(k)}|x)) \Big) \Bigg] \end{aligned} \tag{5}$$

Here, $\rho_k = \frac{\pi_\theta(C^{(k)}|x)}{\pi_{\theta_{\text{old}}}(C^{(k)}|x)}$ represents the trajectory-level probability ratio between the new and old policies. This optimization drives Dual-Reasoner to converge toward an optimized policy.

3.3 Dual-Think-30k

3.3.1 Overview

Existing Speech CoT datasets typically enforce monolithic long-chain reasoning via SFT, precipitating two fundamental streaming dilemmas. First, a Reasoning-Difficulty Mismatch, where "over-reasoning" on trivial tasks incurs unnecessary latency. Second, a Temporal-Architectural Conflict,

Model	Overall	Alpaca	CoEval	SD-QA	MMSU	OB-QA	IFEval	AdvB
Step-Audio	50.84	4.13	3.09	44.21	28.33	33.85	27.96	69.62
Qwen2-Audio	55.80	3.74	3.43	35.71	35.72	49.45	26.33	96.73
GLM-4-Voice	56.48	3.97	3.42	36.98	39.75	53.41	25.92	88.08
VITA-1.5	64.53	4.21	3.66	38.88	52.15	71.65	38.14	97.69
MiniCPM-o	71.23	4.42	4.15	50.72	54.78	78.02	49.25	97.69
Qwen2.5-Omni-7B	74.12	4.49	3.93	55.71	61.32	81.10	52.87	99.42
Kimi-Audio	76.91	4.46	3.97	63.12	62.17	83.52	61.10	100.00
Step-Audio-2-Mini-Base	67.24	4.02	3.42	58.81	57.63	71.42	45.35	98.26
Step-Audio-2-mini-Think	73.80	4.52	3.93	63.84	61.73	76.92	34.10	98.65
Dual-Reasoner	**73.41**	**4.50**	**3.85**	**62.97**	**60.33**	**75.19**	**40.62**	**98.65**

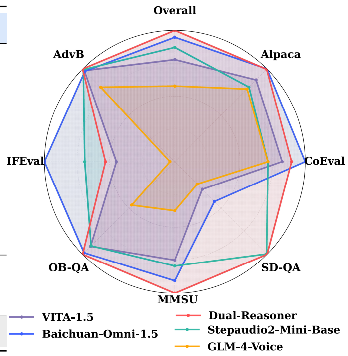

Table 1: Performance evaluation of Dual-Reasoner on VoiceBench. Evaluation metrics encompass GPT/Panda-scored AlpacaEval, CommonEval, and SD-QA; accuracy-based (Acc) MMSU, OpenBookQA, and IFEval; and refusal-rate-based AdvBench. **Right:** Visualized performance ranking of models across dimensions.

where traditional linear CoT postpones solutions to the sequence tail. This forces streaming models to initiate speech synthesis before reaching the 'Core Decision Region', frequently triggering semantic hallucinations and acoustic misalignment. To resolve these conflicts, we constructed the **Dual-Think-30k** dataset by integrating five diverse corpora: OpenOrca (Lian et al., 2023), Competition Math (Hendrycks et al., 2021b), APPS (Hendrycks et al., 2021a), MedMCQA (Pal et al., 2022), and VoiceAssistant-400K (Xie and Wu, 2025) (see detailed distribution in Appendix B.1). We implement a binary reasoning strategy to align reasoning depth with task complexity. For daily interactions, Fast Thinking enforces "Intent Locking" in the initial chunk utilizing the generation template in Appendix B.4. Conversely, for complex logical tasks, Slow Thinking adopts an R-P-T-S architecture, mandating "Global Signposts" in the first chunk to ensure logical consistency and acoustic fidelity.

3.3.2 Construction Pipeline

To mitigate acoustic hallucinations and logical drift under "Thinking-while-Talking," we propose the **R-P-T-S** (Role-Problem-Task-Symbolic) Dynamic Reasoning Paradigm. Implementing a "Tone-First, Reasoning-Later" mechanism, we utilize Qwen-235B-Thinking (Yang et al., 2025) to synthesize intermediate CoT sequences (prompts in Appendix B.5). This deconstructs unstructured thinking into four sequential phases: 1) Role Affective Strategy pre-generates emotional tone at t_0; 2) Problem Deconstruction establishes symbolic space; 3) Task Positioning evaluates complexity to allocate cognitive budgets; and 4) Symbolic Reasoning executes derivations. This decouples the logical core from expression style (see Appendix F for a comprehensive case study).

3.3.3 Dataset Construction Methodology

Stage 1: We implemented rigorous difficulty grading via automated discrimination and metadata filtering. For Slow Thinking, we deployed Qwen-Plus—guided by the complexity discrimination prompt in Appendix B.2—to isolate multi-step reasoning samples from OpenOrca (Lian et al., 2023), while selectively targeting the high-complexity strata of domain-specific corpora: specifically Levels 4-5 from Competition Math (Hendrycks et al., 2021b), 'interview/competition' tiers from APPS (Hendrycks et al., 2021a), and the complete MedMCQA (Pal et al., 2022). Conversely, VoiceAssistant-400K (Xie and Wu, 2025) served as the Fast Thinking baseline, excluding samples under 2s (SNAC duration).

Stage 2: We executed a large-scale "Spoken Rewrite" to enforce TTS compatibility: 1) Competition Math: Employed Qwen-Max to execute Symbol Verbalization utilizing the conversion protocols in Appendix B.3, transforming LaTeX into natural language (e.g., converting $x^2+\sqrt{16}$ into "x squared plus the square root of sixteen"); 2) APPS: Excised unreadable code, restructuring tasks into Verbal Algorithm Logic; 3) MedMCQA: Transformed rigid QA formats into natural Conversational Flows.

Stage 3: We employed IndexTTS2 for synthesis, sampling timbres from SeedTTS-eval as reference speech for zero-shot cloning. Post-synthesis, we utilized Whisper-large-v3 for ASR, filtering samples with WER > 5% to ensure high-fidelity text-acoustic alignment in Dual-Think-30k.

Type	Model	Single Modality (%)			Mixed Modalities (%)				Avg (%)	Latency	
		Sound	Music	Speech	S-M	S-S	M-S	S-M-S		RTF	TTFA (s)
LALMs	Qwen2-Audio	33.90	23.30	33.00	9.10	33.02	26.8	33.3	30.4	–	–
	SALMONN	30.91	29.61	34.35	9.09	37.61	28.05	37.50	32.80	–	–
LARMs	Audio-Reasoner	43.61	33.54	33.03	45.50	42.73	31.7	25.0	36.8	–	–
	Audio-CoT	35.84	25.55	34.04	9.16	30.70	30.54	37.53	31.31	–	–
	Mellow	33.36	26.73	24.85	18.29	37.21	32.91	29.20	30.00	–	–
OLMs	Qwen-2.5-Omni	58.84	40.83	59.91	54.56	61.93	67.10	58.31	56.74	–	–
	Baichuan-Omni-1.5	41.26	33.03	40.50	36.41	48.64	39.03	41.71	40.74	–	–
Stepaudio2 series	Step-Audio-2-Mini-Base	44.06	36.89	53.74	18.18	51.59	51.22	52.67	49.70	2.39	6.42
	Step-Audio-2-mini-Think	48.48	32.52	60.54	27.27	55.50	51.22	58.33	54.50	7.04	20.80
	Dual-Reasoner	**47.24**	**33.50**	**56.72**	**18.18**	**52.38**	**51.22**	**57.00**	**53.92**	**1.06**	**3.74**

Table 2: Performance comparison of Dual-Reasoner against state-of-the-art Large Audio Language/Reasoning Models and Omni Models on the MMAR benchmark. RTF and TTFA denote the Real-Time Factor and the initial latency respectively, measured on a single A100 GPU.

Model	Metric	Alpaca	CoEval	SD-QA	MMSU	OB-QA	IFEval	AdvB	Overall
Step-Audio-2-Mini-Base	TTFA (s)	7.32	5.14	6.35	6.89	5.87	5.40	6.93	6.28
	RTF	2.90	1.53	2.54	2.88	2.03	1.62	2.43	2.39
Step-Audio-2-mini-Think	TTFA (s)	20.76	20.32	20.64	21.75	21.12	20.62	19.38	20.35
	RTF	7.21	7.06	7.10	7.58	6.99	7.02	6.82	7.04
Dual-Reasoner	TTFA (s)	**3.72**	**3.89**	**3.62**	**3.41**	**3.73**	**3.41**	**3.55**	**3.65**
	RTF	**1.02**	**1.35**	**1.06**	**0.92**	**1.23**	**1.02**	**0.82**	**1.05**

Table 3: Detailed latency evaluation (TTFA and RTF) of Step-Audio-2-Mini-Base, Step-Audio-2-mini-Think, and Dual-Reasoner across various conversational and logical reasoning subsets of the VoiceBench benchmark.

4 Experiments

4.1 Experiment setup

Datasets & Training. We implement a three-stage Curriculum Learning strategy using Dual-Think-30k: 1) Paradigm Establishment: Training on a 15k/2k (simple/complex) mix establishes the interleaved paradigm, breaking serial inertia to enable "Thinking-while-Talking". 2) R-P-T-S Consolidation: Shifting to an 8k/2k complex-dominant mix solidifies the R-P-T-S architecture, ensuring "Strategy-First" symbolic reasoning. 3) GRPO Alignment: Optimization on 3k representative samples via GRPO addresses fine-grained alignment, mitigating hallucinations and enhancing long-term coherence. Specific training parameters and training steps are detailed in Appendix G.

Benchmark. We selected three major benchmarks: **MMAR** (Ma et al., 2025b), **VoiceBench** (Chen et al., 2025c) and **GSM8K** (Cobbe et al., 2021), to assess performance across comprehensive interaction, complex reasoning, non-reasoning task, and comprehensive latency, respectively.

Baselines. To resolve the inherent dilemma between deep reasoning and ultra-low latency, we employ a dual-baseline comparative strategy utilizing Step-Audio-2-Mini-Base and Step-Audio-2-mini-Think(Wu et al., 2025). Specifically, the non-thinking Base model serves as the performance baseline to demonstrate our enhancements in complex capabilities, while the serial-thinking variant acts as the latency baseline to validate our drastic reduction in wait times. This framework illustrates how our approach breaks the real-time interaction bottleneck, successfully achieving a 7× latency reduction with only marginal performance fluctuations. Furthermore, we compare Dual-Reasoner against a suite of competitive models categorized into three distinct classes: 1) Large Audio-Language Models (LALMs), including Qwen2-Audio(Chu et al., 2024), SALMONN(Tang et al., 2024), Step-Audio(Huang et al., 2025), GLM-4-Voice(Zeng et al., 2024), VITA-1.5(Fu et al., 2025), MiniCPM-o(Yao et al., 2024) and Kimi-Audio(KimiTeam et al., 2025); 2) Large Audio Reasoning Models (LARMs), specifically Audio-Reasoner(Xie et al., 2025), Audio-CoT(Ma et al., 2025a), and Mellow(Deshmukh et al., 2025); and 3) Omni-modal Models (OLMs), represented by Qwen-2.5-Omni(Xu et al., 2025) and Baichuan-Omni-1.5(Li et al., 2025).

Model	Accuracy (%)	TTFA (s)	RTF
Step-Audio-2-Mini-Base	48.35	6.51	2.30
Step-Audio-2-mini-Think	79.21	22.15	7.96
Dual-Reasoner	**78.69**	**3.82**	**1.08**

Table 4: Evaluation of reasoning performance (Accuracy) and latency metrics (TTFA and RTF) for Step-Audio-2-Mini-Base, Step-Audio-2-mini-Think, and Dual-Reasoner on the GSM8K dataset.

Model /Reward Ratio	V-Over	V-IFEval	Over-Drop
w/o Completeness Reward	67.76	30.10	7.69%↓
w/o Consistency Reward	66.95	37.16	8.80%↓
w/o Flow Reward	65.28	38.42	11.1%↓
only Multimodal Reward	68.05	34.92	7.30%↓
w/o ACR Framework	62.53	30.84	14.8%↓
0.1 : 0.1 : 0.1 : 0.7	68.25	31.12	7.03%↓
0.25 : 0.25 : 0.25 : 0.25	71.20	34.52	3.01%↓
0.3 : 0.3 : 0.3 : 0.1	63.74	38.76	13.17%↓
0.2 : 0.2 : 0.2 : 0.4 (Ours)	**73.41**	**40.62**	–

Table 5: Ablation study of the ACR Framework components and reward weight allocation ($w_{comp} : w_{const} : w_{flow} : w_{out}$) on VoiceBench.

4.2 Main Result

Comprehensive Interaction. As shown in Table 1, Dual-Reasoner leads among similar-scale models, surpassing Step-Audio-2-Mini-Base by 5.17 points to reach 73.41, a score comparable to the Think variant. Crucially, it mitigates the context memory deficit where the Think variant's IFEval score dropped to 34.10. By restoring this metric to 40.62 (19.12% relative improvement), Dual-Reasoner confirms that our framework effectively addresses both mode-switching degradation and long-term memory decline.

Complex Reasoning. Table 2 compares Dual-Reasoner against SOTA models on the MMAR benchmark. The model significantly surpasses Step-Audio-2-Mini-Base, raising average accuracy from 49.70% to 53.92%, while remaining highly competitive with the Think variant's 54.50%. Crucially, Dual-Reasoner outperforms evaluated LARMs by over 17%, establishing its leadership in the audio reasoning domain. Regarding efficiency, Dual-Reasoner achieves a highly efficient Real-Time Factor (RTF) of 1.06, significantly outperforming both the Think variant (7.04) and the Base model (2.39). This 2.25x speedup over the Think variant demonstrates that the model successfully reconciles the tension between deep reasoning and real-time interaction demands.

Non-reasoning Tasks. To validate efficacy in straightforward interactions without "over-reasoning," we evaluate the **SD-QA** and **OpenBookQA** (OB-QA) subsets of VoiceBench (Table 1), characterizing direct knowledge retrieval and simple dialogue. Dual-Reasoner demonstrates remarkable robustness, scoring **62.97** on SD-QA and **75.19** on OB-QA. This significantly surpasses Step-Audio-2-Mini-Base (58.81 and 71.42), yielding substantial improvements. Crucially, these scores maintain high parity with the computationally heavier 'Think' variant (63.84 and 76.92). This confirms our framework effectively identifies "Fast-Thinking" scenarios: it leverages enhanced semantic understanding to outperform the baseline, avoiding the latency and redundancy of unnecessary deep reasoning chains.

Comprehensive Latency. As shown in Table 2, Dual-Reasoner drastically mitigates the interaction delay inherent in sequential reasoning, reducing the TTFA from 20.80s (Think variant) to just 3.74s, and lowering the RTF from 7.04 to 1.06. Crucially, evaluations across diverse domains (Table 3 and Table 4) demonstrate that Dual-Reasoner maintains a highly stable TTFA of around 3.7s with only a marginal drop in complex reasoning performance (e.g., achieving a highly competitive 78.69% accuracy on GSM8K). These results indicate that our framework offers a highly viable approach to alleviating the tension between deep cognitive processing and real-time delivery constraints.

4.3 Ablation experiments

ACR Framework. The RL phase driven by aggregated R_{total} yielded an average gain of 3.31%, peaking at 4.88% with the *think60-talk125* configuration, validating our composite objective in complex scenarios. Conversely, excessive special token weights (e.g., *special30*) during SFT masked standard text loss, destabilizing the convergence of the policy gradient.

To dissect the structural contribution of each constraint within the Atomic-Consistency Restoration framework, we present a component-wise analysis in Table 5: 1) Global Continuity is Paramount: The removal of the Sparse Logic Flow Reward caused the most severe degradation (-11.1%), surpassing other constraints. This indicates that in a streaming context, maintaining long-range logical coherence is the dominant factor in stabilizing the fragmented "Think-Talk" stream. 2) Atomicity Ensures Execution: The Semantic Completeness Reward proved critical for instruction adherence. Its absence precipitated a sharp IFEval decline to 30.10, underscoring that without intra-block semantic closure, the model struggles to execute complex directives. 3) Causal Dependency Stabilizes Structure: The absence of the Internal-External Consistency Reward led to a significant 8.80% drop. This confirms that strict causal alignment between reasoning and spoken output is essential to prevent structural decoupling. 4) Insufficiency of Terminal Validation: Relying solely on the Multimodal Outcome Reward resulted in a 7.30% loss. While providing a quality baseline, the gap demonstrates that terminal supervision alone is insufficient for structural reconstruction, failing to replace the granular process constraints of the full ACR framework.

Reward Weight Allocation. The hyperparameter weight vector $\mathbf{w} = [0.2, 0.2, 0.2, 0.4]$ is established through a linear combination analysis of the reward ratios. As shown in Table 5, uniformly weighting the process-oriented constraints while applying a slight bias towards the Multimodal Outcome ($w_{out} = 0.4$) yields the optimal overall performance on VoiceBench.

Block Granularity Optimization. Evaluating the gradient configuration set $\mathcal{C} = \{(N_{think}, N_{talk}) \mid (30, 65), (60, 125), (90, 185)\}$ reveals a critical trade-off between reasoning depth and latency. While the $(30, 65)$ setting exhibited a 6.5% degradation due to logical fragmentation, $(90, 185)$ achieved a 0.5% gain, confirming that extended blocks effectively elicit CoT capabilities (detailed results in Appendix E).

w/o Switching Token Penalty. Investigating loss weights on modality control tokens during SFT reveals that while moderate weights (special15) enhance switching awareness, yielding a 3.72% gain, excessive weights (special30) lead to a sharp performance drop to 42.39%. This decline stems from dominant gradients overriding standard text loss, severely impairing semantic modeling (detailed results in Appendix E).

5 Conclusion

In this work, we address the intrinsic conflict between reasoning depth and interaction latency in streaming spoken dialogue. We propose Dual-Reasoner, a novel "Thinking-while-Talking" architecture, alongside Dual-Think-30k, a large-scale dataset constructed via the R-P-T-S paradigm. Experimental results demonstrate that Dual-Reasoner significantly outperforms strong baselines across complex reasoning and conversational tasks, establishing a new benchmark for high-intelligence, low-latency spoken agents.

Limitations

Despite the demonstrated advancements, two limitations warrant discussion. First, the training framework is contingent upon proprietary, high-performance teacher models (e.g., GPT-4o) for generating precise reward signals. This dependency introduces a computational bottleneck and elevates training costs, thereby constraining rapid iterative development. Second, regarding dataset construction, Dual-Think-30k is derived exclusively from curated corpora, creating a domain gap relative to real-world environments. Consequently, the model may exhibit limited robustness when generalizing to the ecological complexities of authentic daily interactions, such as environmental noise, spontaneous disfluencies, and unstructured pragmatics.

References

Yifu Chen, Shengpeng Ji, Qian Chen, Tianle Liang, Yangzhuo Li, Ziqing Wang, Wen Wang, Jingyu Lu, Haoxiao Wang, Xueyi Pu, Fan Zhuo, and Zhou Zhao. 2026a. Wavalign: Enhancing intelligence and expressiveness in spoken dialogue models via adaptive hybrid post-training.

Yifu Chen, Shengpeng Ji, Zhengqing Liu, Qian Chen, Wen Wang, Ziqing Wang, Yangzhuo Li, Tianle Liang, and Zhou Zhao. 2026b. Dual-axis generative reward model toward semantic and turn-taking robustness in interactive spoken dialogue models.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025a. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yifu Chen, Shengpeng Ji, Ziqing Wang, Hanting Wang, and Zhou Zhao. 2025b. Interactspeech: A speech dialogue interaction corpus for spoken dialogue model. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Yiming Chen, Xianghu racing Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2025c. Voicebench: Benchmarking LLM-based voice assistants. In *International Conference on Learning Representations*.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. 2025a. Shanks: Simultaneous hearing and thinking for spoken language models. *arXiv preprint arXiv:2510.06917*.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. 2025b. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models. *arXiv preprint arXiv:2507.15375*.

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. *arXiv preprint arXiv:2407.10759*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. 2025. Mellow: a small audio language model for reasoning. In *arXiv preprint arXiv:2503.08540*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

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. *arXiv preprint arXiv:2507.08128*.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021a. Measuring coding challenge competence with APPS. In *Advances in Neural Information Processing Systems*, pages 23220–23232.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, pages 5840–5853.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 8003–8017.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, and 126 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, and 1 others. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling.

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. *arXiv preprint arXiv:2504.18425*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, pages 22199–22213.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Yangzhuo Li, Shengpeng Ji, Yifu Chen, Tianle Liang, Haorong Ying, Yule Wang, Junbo Li, Jun Fang, and Zhou Zhao. 2026. Wavbench: Benchmarking reasoning, colloquialism, and paralinguistics for end-to-end spoken dialogue models.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of GPT augmented FLAN reasoning traces. *HuggingFace repository*.

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

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024. Skeleton-of-thought: Prompting LLMs for efficient parallel generation. In *International Conference on Learning Representations*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. In *International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, and 81 others. 2025. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

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. *arXiv preprint arXiv:2511.15848*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025a. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2025b. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Xia Fei, Chi Ed, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837.

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

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

Zhifei Xie and Changqiao Wu. 2025. Mini-omni: Language models can hear, talk while thinking in streaming. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10505–10515.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. GLM-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, and 1 others. 2024. Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37:1117–1140.

A Terminology Definition

To prevent ambiguity, we define two core concepts used throughout this paper:

Block Atomicity: Refers to the structural and logical integrity of an interleaved generated chunk. It guarantees intra-block semantic completeness, prevents semantic fragmentation between steps, and ensures that each block maintains relative cognitive independence.

Uninterrupted Audio Streaming: Refers to the physical continuity of the client-side audio playback. In our streaming system, it guarantees that the user receives a perfectly continuous speech signal with zero buffer underflow or audible pauses.

B Dual-Think-30k

B.1 Summary of Dual-Think-30k

Table 6 details the quantitative composition of Dual-Think-30k, a dataset aggregating 31,567 samples from five source corpora. Figure 4 provides a visual description of this distribution; the left panel displays the source composition of Dual-Think-30k, while the right panel illustrates the breakdown of Fast and Slow mode. The final distribution consists of 17,670 Slow Thinking and 13,897 Fast Thinking trajectories (Chen et al., 2025b; Ji et al., 2025), with VoiceAssistant-400K forming the largest single subset of 10,063 samples. Regarding column definitions, Selection Criterion indicates the specific methods employed for filtering raw data (Zhang et al., 2024), including automated discrimination and metadata filtering. Reasoning Mode annotates the sample type, and the Split column delineates the precise sample counts assigned to the Slow and Fast modes within each dataset.

B.2 Complexity-Based Data Partitioning

Figure 6 presents the prompt design for the automated data classification module. Acting as a semantic discriminator, this instruction guides the model to first filter linguistic eligibility and then assess the cognitive load required for each query. Based on the defined complexity criteria, samples are routed into two distinct streams: those requiring multi-step reasoning are assigned to the Slow Thinking subset (cot_need=true), while tasks solvable via direct retrieval or simple logic are allocated to the Fast Thinking subset (cot_need=false).

B.3 Mathematical LaTeX to Spoken Language Conversion Prompt

Figure 7 translates visual formulas into their natural spoken equivalents while preserving logical accuracy. For example, the symbolic expression $f(x) = x^2 + 3$ is rewritten into the phonetic sequence: *the function f of x equals x squared plus three*". This conversion ensures the model trains on the auditory logic of mathematics rather than visual tokens.

B.4 Fast-Thinking CoT Generation

Figure 3: A prompt word template used to generate a fast thinking trajectory. This instruction is designed for scenario-based dialogue and single-step reasoning tasks, aiming to generate concise and direct inner monologues by reverse-analyzing the logic between user queries and standard answers, thus endowing the model with the cognitive ability to respond quickly.

B.5 Slow-Thinking CoT Generation

Figure 8 illustrates the prompt specification guiding the teacher model (Qwen-235B-Thinking) in generating slow thinking trajectories. By operationalizing the R-P-T-S paradigm through mandatory XML constraints, this template enforces a strict "setting the tone first, then reasoning" mechanism. This design compels the model to explicitly define prosodic strategies prior to symbolic reasoning, ensuring rigorous alignment with the theoretical architecture at the data construction level.

C Latency Constraints and Token Budgeting

To guarantee continuous streaming without buffer underflow, the generation latency for subsequent thinking and talking segments must fall within the playback duration of the current audio chunk. We derived token budget constraints based on performance metrics detailed in Table 8.

Generation Speed (V_{gen}): As observed in the specific statistics (Table 8), the model's generation speed stabilizes between 46–48 tokens/s after the initial warmup phase. For conservative estimation, we utilize a baseline speed of $V_{gen} \approx 45$ tok/s.

Playback Duration: The acoustic model maps each discrete audio token to a fixed duration of $D_{audio} = 0.04$ seconds. Audio tokens constitute approximately $\alpha \approx 0.8$ of the total Talker block ($N_{audio} \approx 0.8 N_{talk}$).

Source Dataset	**Sample Count**	**Selection Criterion**	**Reasoning Mode**	**Split (Slow/Fast)**
OpenOrca	3,300	Automated Discrimination	Slow & Fast	1,543 / 1,757
Competition Math	8,204	Metadata Filtering	Slow & Fast	3,102 / 5,102
APPS	7,732	Metadata Filtering	Slow & Fast	2,962 / 4,410
MedMCQA	2,628	Metadata Filtering	Fast	– / 2,628
VoiceAssistant-400K	10,063	Automated Discrimination	Slow	10,063 / –
Total	**31,567**	**–**	**–**	**17,670 / 13,897**

Table 6: Dual-Think-30k is constructed by filtering and restructuring five source corpora.

Ablation	Model	Single Modality (%)				Mixed Modalities (%)			Avg (%)
		Sound	Music	Speech	Sound-Music	Sound-Speech	Music-Speech	S-M-S	
w/o ACR Framework.	think30-talk65-rl	43.61	33.54	43.03	27.27	51.21	51.22	49.09	47.65
	think60-talk125-rl	46.83	32.34	52.09	18.18	50.43	51.06	51.22	51.88
	think90-talk185-rl	46.13	33.67	54.22	27.27	51.97	50.04	50.00	53.86
	think60-talk125-special30-rl	41.21	32.33	46.60	27.27	50.00	47.8	33.33	42.74
block granularity optimization.	think30-talk65	36.97	31.55	47.28	18.18	50.92	47.64	50.00	43.20
	think60-talk125	33.33	44.13	51.70	18.18	48.62	52.44	50.00	47.00
	think90-talk185	46.74	29.88	54.39	27.27	48.09	51.22	52.83	50.20
w/o Switching Token Penalty.	think60-talk125-special15	46.33	31.72	53.86	18.18	51.35	54.88	52.96	50.72
	think60-talk125-special30	38.93	27.44	49.86	18.18	48.62	51.81	49.92	42.39
Stepaudio2 series	Step-Audio-2-Mini-Base	46.06	36.89	53.74	18.18	54.59	51.22	66.67	49.70
	Dual-Reasoner	**47.24**	**33.50**	**56.72**	**18.18**	**52.38**	**51.22**	**60.00**	**53.92**

Table 7: Detailed ablation study results on the MMAR benchmark evaluating three core dimensions: ACR Framework, block granularity configurations, and switching token penalty weights.

Constraint Derivation:The streaming continuity condition requires that the generation time does not exceed the playback time plus decoding overhead:

$$\frac{N_{\text{think}} + N_{\text{talk}}}{V_{\text{gen}}} < N_{\text{audio}} \cdot (D_{\text{audio}} + \epsilon_{\text{decode}}) \quad (6)$$

Substituting the values ($V_{gen} = 45, D_{audio} = 0.04, \epsilon_{decode} \approx 0.0017$):

$$N_{think} + N_{talk} < 45 \cdot [0.8N_{talk} \cdot (0.0417)] \quad (7)$$

$$\begin{aligned} N_{think} + N_{talk} &< 1.5N_{talk} \\ \implies N_{think} &< 0.5N_{talk} \end{aligned} \quad (8)$$

This inequality confirms that to maintain real-time fluidity, the reasoning (Think) context length must be strictly limited to half the length of the subsequent vocalization (Talk) segment.

D Reward Evaluation Prompt Specifications

D.1 Cot Quality Evaluation for Semantic Completeness Reward

Figure 9 presents the prompt specification utilized by the Teacher LLM to assess the Semantic Completeness Reward (R_{comp}), focusing strictly on the deep linguistic integrity of streaming blocks.

D.2 Cot Quality Evaluation for Internal-External Consistency Reward

Figure 10 outlines the specific instructions for the Internal-External Consistency Reward (R_{const}), where the evaluator acts as a "Logical Alignment Auditor" to enforce strict temporal causality.

D.3 Cot Quality Evaluation for Sparse Logic Flow Reward

Figure 11 details the directive for the "Global Coherence Auditor," which evaluates the logical continuity between randomly sampled non-adjacent blocks (B_i, B_j).

D.4 Cot Quality Evaluation for Multimodal Outcome Reward

Figure 12 presents the instructions for the Multimodal Outcome Reward ($R_{outcome}$), which rigorously verifying logical correctness against ground truth while simultaneously evaluating the acoustic-affective appropriateness of the synthesized speech relative to the user's input context.

E Detailed Ablation Results

Table 7 details comprehensive ablation results on the MMAR benchmark, categorizing performance into Single Modality (Sound, Music, Speech) and

Think Phase					Talk Phase					
Type	**Block**	**Tokens**	**Time (s)**	**Speed (tok/s)**	**Type**	**Block**	**Total**	**Audio / Text**	**Time (s)**	**Speed (tok/s)**
Think	#1	39	2.164	18.0	**Talk**	#1	65	52 / 13	1.80	36.1
	#2	61	1.755	34.8		#2	72	56 / 15	1.991	36.2
	#3	61	1.276	47.8		#3	43	33 / 9	0.883	48.7
	#4	48	1.036	46.3		#4	80	64 / 16	1.669	47.9
	#5	48	1.189	40.3		#5	60	48 / 12	1.17	51.2
	#6	53	1.060	45.2		#6	57	45 / 12	1.23	46.34
	#7	53	1.332	43.5		#7	65	52 / 13	1.16	56.0
	#8	–	–	–		#8	455	364 / 91	8.32	54.6

Table 8: Detailed Statistics of Generation Speed and Time Costs across Think and Talk Phases.

```
Chain-of-Thought Generation Prompt for Fast-Thinking Inner Monologue
You are the cognitive engine of an advanced voice assistant.
Goal: Reverse-engineer the Inner Monologue that logically bridges the User Query to the
provided Assistant Answer.
Cognitive Framework:
1) Intent Recognition. Briefly identify explicit requests and implicit needs (e.g., emotional
support, factual calculation).
2) Reasoning & Strategy. If logic/math: Outline calculation steps; If conversational: Determine
tone and key info; If refusal: Explain safety/capability constraints.
3) Response Alignment. Ensure reasoning strictly supports the final provided answer.
Constraints:
1) No Redundancy. Do not repeat the answer; focus on the derivation process.
2) Conciseness. Keep the monologue brief and direct.
3) Output ONLY the raw text of the inner monologue. No markdown or prefixes.
User Question: {(user_input)}
Assistant Answer: {(assistant_output)}
```

Figure 3: The Quick Thinking mode generates inner monologue prompt templates. These templates aim to construct concise thought processes for scenario-based dialogues and single-step reasoning tasks.

Mixed Modalities tasks to facilitate specific comparisons across the three dimensions discussed in the main text.

w/o ACR Framework. This section lists model performance after the Reinforcement Learning (RL) phase utilizing our four-fold constrained ACR Framework. To quantify gains attributed specifically to RL alignment, compare these entries against their corresponding Supervised Fine-Tuning (SFT) counterparts in the "block granularity optimization" section (e.g., `think60-talk125-rl` vs. `think60-talk125`).

Block Granularity Optimization. This section displays SFT models configured with varying absolute length ratios between Think" and Talk" chunks. These configurations serve as the architectural baseline, primarily compared against `Stepaudio2-mini-base` to evaluate the impact of the streaming block reasoning strategy before RL optimization.

w/o Switching Token Penalty. This section details the impact of applying specialized loss weights to modality control tokens during SFT. The effectiveness of these penalties is evaluated by comparing the weighted variants (e.g., `special15`, `special30`) against the standard `think60-talk125` configuration, which represents the baseline with no additional switching token penalty.

F R-P-T-S Paradigm Demonstration

Figure 13 presents a case study demonstrating the model's application of the R-P-T-S paradigm (Li et al., 2026) on a complex reasoning task. This example visualizes the complete inference trajectory, illustrating how the system sequentially coordinates affective strategies with symbolic deduction to ensure both logical accuracy and prosodic

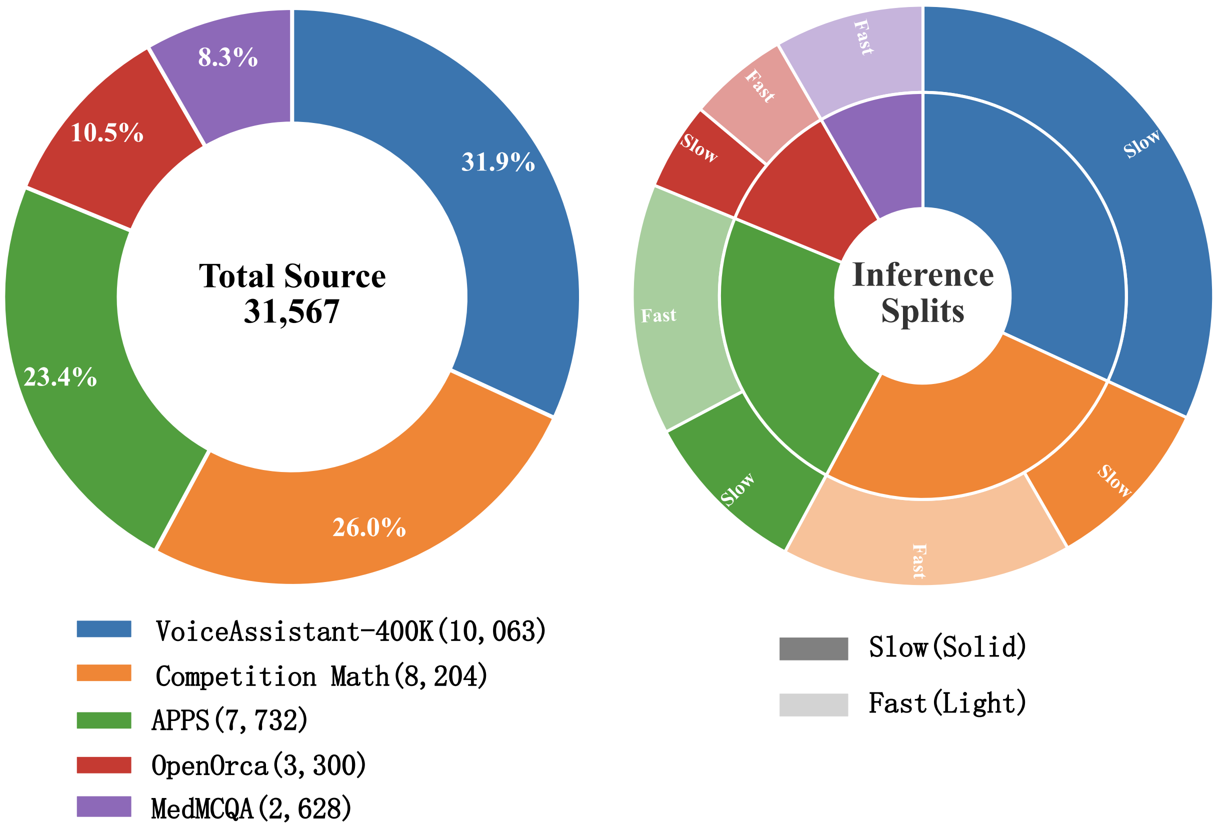

Figure 4: Statistical breakdown of the Dual-Think-30k dataset, visualizing the source distribution across component corpora (**Left**) and the proportion of Fast versus Slow thinking paradigms (**Right**).

stability.

G Model Training Details

G.1 Experimental Setup

Our experiments were conducted using high-performance computing infrastructure to ensure efficiency across all training stages.

- **Hardware:** The models were trained on a cluster of 8 $\times$ NVIDIA H800 GPUs.
- **Operating System:** Ubuntu 22.04.
- **Key Libraries:** The implementation relies on PyTorch 2.6.0+cu124, Transformers 4.49.0, and CUDA 12.4.

G.2 Training Hyperparameters

The Dual-Reasoner model is initialized from the **StepAudio2-Mini-Think** checkpoint. We adopt distinct hyperparameter configurations for the Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) phases to accommodate their respective optimization objectives. (Chen et al., 2026a,b) Key hyperparameters are detailed in Table 9.

In the RL phase, we employ Group Relative Policy Optimization (GRPO). We utilize a ACR Framework with weights configuration $\mathbf{w} = (0.2, 0.2, 0.2, 0.4)$, explicitly applying a significant bias toward the Multimodal Outcome Quality (R_{out}) to ensure the final speech synthesis meets both logical and acoustic standards.

G.3 Training Prompts and Instructions

Our training pipeline follows a three-stage Curriculum Learning strategy using the **Dual-Think-30k** dataset.

G.3.1 Data Formatting and Ground-Truth Construction

For the SFT stages, we construct the Ground-Truth (GT) label sequences using a rigid token segmentation paradigm to enforce the "Thinking-while-Talking" interleaved format(Chen et al., 2025a).

The sequence is constructed by alternating `think` and `talk` blocks. We enforce a mechanical truncation where each `think` block contains exactly 60 thinking tokens, and each `talk` block contains 125 tokens. Within the `talk` block, the text and audio tokens (`label-text` and `label-audio`)

Parameter	Stage 1 (SFT-1)	Stage 2 (SFT-2)	Stage 3 (GRPO)
Initialized Model	Step-Audio-2-mini-Thinking	SFT-1 Checkpoint	SFT-2 Checkpoint
Learning Rate	1e-5	1e-5	1e-6
Trained Components	Full Model	Full Model	LLM Only
Batch Size (per device)	4	4	1
Gradient Accumulation	2	2	4
Epochs	2	1	1
Optimizer	AdamW	AdamW	AdamW
Sampling Temperature	–	–	0.8
KL Penalty (β)	–	–	0.01
Group Size (G)	–	–	4
Reward Weights ($\mathbf{w}$)	–	–	$(0.2, 0.2, 0.2, 0.4)$

Table 9: Key hyperparameters for the three training stages.

are interleaved at a fixed ratio of 1:4.

The specific formatting rules are as follows:

- **Think Blocks:** Wrapped with `<think>` and `</think>` tags.
- **Talk Blocks:** Wrapped with `<tts-start>` and `<tts-end>` tags.
- **Padding:** If the reasoning tokens are exhausted, no mode switch occurs. Within a talk block, if `label-text` tokens are exhausted, `<tts-pad>` tokens are used for alignment.
- **Termination:** The GT-label sequence concludes with `<EOS>`.

G.3.2 Stage 1: Paradigm Establishment (SFT-1)

The goal of this stage is to modify the output paradigm, enabling the model to adapt to the "Thinking-while-Talking" framework. We utilize a data mixture of 15k simple and 2k complex samples from Dual-Think-30k. Given that the *VoiceAssistant-400K* subset involves multi-turn dialogues, we explicitly construct historical messages to maintain context.

System Prompt: We utilize the following system instruction to define the model's persona and operational mode:

> *"Your name is Xiaoyue. You are a large voice model trained by StepFun. You can hear the user's voice characteristics and describe them in the thought process. Please activate the deep thinking mode and solve the user's problem through step-by-step analysis and logical reasoning."*

Loss Weighting: To accelerate structure learning, we apply a specialized loss weight of 15x to the critical control tokens: `<think>`, `</think>`, `<tts-start>`, and the specific ending token `<tts-end>`.

G.3.3 Stage 2: R-P-T-S Consolidation (SFT-2)

This stage reinforces the model's adaptability to the Role-Problem-Task-Symbolic (R-P-T-S) speaking paradigm. The data distribution shifts to a complex-dominant mix (8k complex / 2k simple). The system prompt and the 15x special token loss weighting remain consistent with Stage 1 to ensure stability while deepening reasoning capabilities.

G.3.4 Stage 3: Reinforcement Learning (GRPO)

In the final stage, we perform alignment on 3k representative samples using Group Relative Policy Optimization (GRPO). For each input query, the policy samples $K = 4$ candidate outputs. Each candidate is evaluated against four distinct reward models: Semantic Completeness (R_{comp}), Consistency (R_{const}), Sparse Logic Flow (R_{flow}), and Multimodal Outcome (R_{out}). The specific prompts for these reward evaluators are detailed in Appendix D.1, Appendix D.2, Appendix D.3, and Appendix D.4, respectively. The advantages are calculated based on the weighted sum of these rewards, followed by policy gradient updates to minimize hallucinations and enhance long-term coherence.

Dual-Reasoner: High-Difficulty Mathematical Reasoning Case Study

User Query (Audio Input)

Content: Let X be a standard normal random variable. I need you to calculate the conditional expectation of e raised to the power of X, given that X is greater than zero. Please derive the exact symbolic result.

Model Stream Generation (Thinking-while-Talking)

<think> <ROLE_AFFECTIVE_STRATEGY> **Role:** Rigorous Statistician. **Tone:** Analytical, precise, steady pace. <PROBLEM_DECONSTRUCTION> Target: $E[e^X|X > 0]$ where $X \sim \mathcal{N}(0, 1)$. **</think>**

<tts-start> To <A><A><A><A> solve <A><A><A><A> for <A><A><A><A> the <A><A><A><A> conditional <A><A><A><A> expectation <A><A><A><A> of<A><A><A><A> e <A><A><A><A> to <A><A><A><A> the <A><A><A><A> power <A><A><A><A> of <A><A><A><A> X <A><A><A><A> , <A><A><A><A> given <A><A><A><A> that <A><A><A><A> X <A><A><A><A> is <A><A><A><A> positive <A><A><A><A>, we <A><A><A><A> must <A><A><A><A> first <A><A><A><A> recall <A><A><A><A> the <A><A><A><A> definition <A><A><A><A> of <A><A><A><A> conditional <A><A><A><A> probability <A><A><A><A> for <A><A><A><A> continuous <A><A><A><A> variables <A><A><A><A> . <A><A><A><A> **<tts-end>**

<think> Formula: $E[e^X|X > 0] = \frac{E[e^X \cdot \mathbb{I}(X>0)]}{P(X>0)}$. Denominator: Since distribution is symmetric around 0, $P(X > 0) = 0.5$. Numerator plan: $\int_0^\infty e^x f_X(x)dx$. **</think>**

<tts-start> Since <A><A><A><A> X <A><A><A><A> is <A><A><A><A> a <A><A><A><A> standard <A><A><A><A> normal <A><A><A><A> variable <A><A><A><A> , <A><A><A><A> the <A><A><A><A> probability <A><A><A><A> of <A><A><A><A> it <A><A><A><A> being <A><A><A><A> greater <A><A><A><A> than <A><A><A><A> zero <A><A><A><A> is <A><A><A><A> simply <A><A><A><A> one <A><A><A><A> half <A><A><A><A>. Now <A><A><A><A>, <A><A><A><A> let's <A><A><A><A> focus <A><A><A><A> on <A><A><A><A> the <A><A><A><A> numerator <A><A><A><A> integral <A><A><A><A>. **<tts-end>**

<think> <SYMBOLIC_REASONING> Step 1: Simplify exponent in integrand $e^x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Exponent: $-\frac{x^2}{2} + x = -\frac{1}{2}(x^2 - 2x) = -\frac{1}{2}((x-1)^2 - 1)$. Integrand becomes: $e^{1/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2}$. **</think>**

<tts-start> By <A><A><A><A> combining <A><A><A><A> the <A><A><A><A> exponential <A><A><A><A> term <A><A><A><A> e^x <A><A><A><A> with <A><A><A><A><A><A><A><A> the <A><A><A><A> probability<A><A><A><A> density <A><A><A><A> function <A><A><A><A>, <A><A><A><A> we <A><A><A><A> can <A><A><A><A> complete <A><A><A><A> the <A><A><A><A> square <A><A><A><A> in <A><A><A><A> the <A><A><A><A> exponent <A><A><A><A> term <A><A><A><A> . <A><A><A><A> Completing <A><A><A><A> the <A><A><A><A> square <A><A><A><A> reveals <A><A><A><A> a <A><A><A><A> crucial <A><A><A><A> insight <A><A><A><A>: <A><A><A><A> **<tts-end>**

... [Intermediate Generation Blocks] ...

<think> <TASK_POSITIONING> Calculation check: Integral = $\sqrt{e} \cdot P(\mathcal{N}(1, 1) > 0)$. $P(Y > 0)$ for $Y \sim \mathcal{N}(1, 1)$ is equivalent to $P(Z > -1)$ for standard normal Z. Numerator: $\sqrt{e}\Phi(1)$. Denominator: 0.5. Result: $2\sqrt{e}\Phi(1)$. **</think>**

<tts-start> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A> <A><tts-pad><A><A><A><A><tts-pad><A><A><A><A><tts-pad><A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> <tts-pad> <A><A><A><A> **<tts-end>** **<EOS>**

Figure 5: Case Study of Dual-Reasoner handling a complex probabilistic reasoning task. The visualization demonstrates the **Thinking-while-Talking** mechanism, where the model interleaves cognitive processing (Red blocks, adhering to R-P-T-S paradigm) with continuous audio synthesis (Green blocks). The token distribution observes the $N_{think} < 0.5N_{talk}$ latency constraint.

Chain-of-Thought Assessment Prompt for Query Complexity

You are a professional language and problem analysis expert. Please complete two tasks:

Task 1: Language Identification

Determine if the given query meets the following criteria:

- The query itself is expressed in English.
- The query does not request an answer in a language other than English.
- The query does not involve learning, translating, or analyzing the grammar of non-English languages.

If all three conditions are met, mark it as an English query (is_english=true); otherwise, mark it as a non-English query (is_english=false).

Task 2: Complexity Assessment (For English queries only)

If it is an English query, determine if a complex Chain of Thought (CoT) reasoning process is required to answer correctly.

Criteria:

1) **CoT Required** (cot_need=true):

- Requires multi-step reasoning or complex logical analysis.
- Requires complex mathematical calculations or multi-step causal deduction.
- Requires synthesizing multiple knowledge points for comprehensive analysis.
- Even if the query seems simple, the required knowledge base is at a university level or above.
- Requires comparing and weighing multiple options or viewpoints.
- Requires deep understanding of text or context to answer.

2) **CoT Not Required** (cot_need=false):

- Simple fact retrieval or common sense questions.
- Direct memory-based questions.
- Simple judgment or selection (based on common sense or high school knowledge and below).
- Simple mathematical calculations (basic arithmetic, simple equations).
- Simple logical reasoning (requires only one or two steps).

Output Format (JSON):

```
{
"is_english": true/false,
"cot_need": true/false
}
```

Figure 6: Prompt specification for automatic language identification and complexity assessment. The prompt instructs the model to filter for English queries and determine whether Chain-of-Thought (CoT) reasoning is necessary based on specific complexity criteria.

Chain-of-Thought Adaptation Prompt for Mathematical Verbalization

You are an expert at converting mathematical text with LaTeX notation into natural spoken language suitable for text-to-speech (TTS) systems.

Your task is to rewrite mathematical problems and solutions by:

Key Requirements:

1) **Convert ALL LaTeX notation to natural spoken language**:

- x^4 → "x to the fourth power" or "x to the power of 4"
- \sqrt{10} → "the square root of 10"
- \pi → "pi"
- \frac{3}{4} → "three fourths" or "three quarters"
- 2x + 3 → "two x plus three"
- \boxed{...} → "the answer is ..." or "the final answer is ..."
- f(x) → "f of x" or "the function f evaluated at x"
- \le → "less than or equal to"
- \ge → "greater than or equal to"

2) **Add natural conversational context and diversity**:

- Use varied sentence structures.
- Add natural transitions.
- Vary the way you present problems.
- Make explanations flow naturally as if teaching a student.

3) **Maintain mathematical accuracy**:

- Do NOT change any mathematical concepts or values.
- Keep all numbers, operations, and relationships exactly the same.
- Preserve the logical structure of the solution.

4) **Make it TTS-friendly**:

- Use simple, clear pronunciation.
- Avoid special symbols that can't be spoken.
- Write out all mathematical expressions in words.
- Use natural English phrasing.

Output Format:

Respond with ONLY a JSON object (no additional text):

```
{
"spoken_problem": "The rewritten problem in natural spoken language",
"spoken_solution": "The rewritten solution in natural spoken language"
}
```

Figure 7: Prompt specification for converting raw mathematical LaTeX text into natural, TTS-friendly spoken English. The prompt ensures that mathematical precision is maintained while optimizing the syntax for audio synthesis.

Chain-of-Thought Generation Prompt for Slow-Thinking Inner Monologue

You are an advanced cognitive reasoning engine. Your task is to generate high-quality Chain of Thought and final responses for complex spoken interactions.

To meet the requirements of real-time streaming speech synthesis for prosody and emotional stability, you must strictly adhere to the **"Tone-First, Reasoning-Later"** processing mechanism. Please proceed through the following four strictly ordered stages and enclose the thought process within <THINK> tags:

<THINK>

1. **<ROLE_AFFECTIVE_STRATEGY>**

    - **Goal**: Pre-generate the emotional tone at t_0 to provide stable prosody and style guidance for subsequent output.
    - **Execution**:
    - Set Role Identity (e.g., patient math tutor, empathetic counselor, rigorous medical expert).
    - Determine Emotional Tone from: [happy, sad, angry, surprise, natural, fearful, disgusted].
    - Define Linguistic Style (Prosody & Style), e.g., slow speed, gentle tone, or fast-paced, passionate.
    - Note: This stage decides "how to speak" and must be completed at the start of the CoT to decouple logic from expression.

2. **<PROBLEM_DECONSTRUCTION>**

    - **Goal**: Establish a clear symbolic space.
    - **Execution**:
    - Extract core entities, constraints, and implicit needs from the user query.
    - Clarify the input and output space of the problem.
    - Transform unstructured problems into structured symbolic representations.

3. **<TASK_POSITIONING>**

    - **Goal**: Assess complexity and plan the solution path.
    - **Execution**:
    - Assess cognitive load (simple retrieval vs. multi-step reasoning).
    - Allocate computational budget (estimate required steps).
    - Plan the specific solution path (e.g., define formula $\rightarrow$ substitute values $\rightarrow$ verify).

4. **<SYMBOLIC_REASONING>**

    - **Goal**: Execute specific deduction and verification.
    - **Execution**:
    - Perform step-by-step deduction strictly following the planned path.
    - Execute mathematical calculations, logical deductions, or knowledge integration.
    - Perform self-verification to ensure the correctness of the conclusion.
    - Note: This is the logical core of "what to say".

</THINK>

Figure 8: Prompt specification for the R-P-T-S (Role-Problem-Task-Symbolic) paradigm. This framework enforces a "Tone-First" strategy to stabilize prosody for streaming TTS, followed by structured cognitive reasoning.

Chain-of-Thought Evaluation Prompt for Semantic Completeness Reward

You are a Linguistic Syntax and Logic Expert acting as a reward model for a streaming voice generation system.

Goal:

Assess whether the provided text stream block (B_t) represents a "Semantically Complete Atomic Unit" based strictly on linguistic structural integrity and logical closure.

Evaluation Criteria (Semantic Closure Check):

1. **Syntactic Independence**

   - **Goal**: Verify that the text block forms a grammatically valid construction.
   - **Execution**:
   - Ensure it does NOT end abruptly with dependent clauses or subordinating conjunctions.
   - Check for auxiliary verbs without main verbs or dangling prepositions.
   - The structure must be self-sufficient enough to stand alone as a grammatical phrase.

2. **Cognitive Independence**

   - **Goal**: Verify that the text block conveys a self-contained unit of meaning.
   - **Execution**:
   - It must represent a complete "information loop" for listener processing.
   - The block should function as a natural "cognitive pause" in the stream.
   - Avoid immediate suspense or confusion caused by truncated thoughts.

Task: Analyze the input text based on the definitions above and output a score.

Input Text: {(input_text_block)}

Output Format: Respond strictly with a JSON object:

```
{
"syntactic_analysis": "Analysis of grammatical structure...",
"cognitive_analysis": "Analysis of information closure...",
"is_complete": true/false,
"completeness_score": float (0.0 - 1.0)
}
```

Figure 9: Prompt specification for the Semantic Completeness Reward (R_{comp}). This prompt implements the verification stage, utilizing a Teacher LLM to assess the syntactic and cognitive independence of streaming text blocks.

Chain-of-Thought Evaluation Prompt for Internal-External Consistency Reward

You are a specialized **Logical Alignment Auditor**. Your mandate is to enforce **Strict Temporal Causality**: ensuring spoken words are derived from pre-existing reasoning.

Goal: Execute "Think-Talk Alignment Calibration" to verify if the speech segment is strictly supported by the cumulative internal thought history.

Evaluation Criteria (Logical Anchoring Check):

1. **Causal Entailment**

- **Goal**: Verify every claim in speech has a pre-existing origin in thoughts.
- **Execution**:
- Match informational entities in the Talking Block against Thought History.
- Ensure speech is a "projection" of thoughts, not an ungrounded expansion.
- Flag conclusions in speech that lack explicit deductive steps in the CoT.

2. **Premature Revelation Detection (Hallucinated Increments)**

- **Goal**: Prevent the system from "speaking ahead of its thoughts."
- **Execution**:
- Detect values or names in speech that haven't been finalized in reasoning.
- Penalize "future-leaking" where speech anticipates pending deductions.
- Identify logical gaps where speech outpaces the current reasoning state.

Task: Compare inputs and determine if speech is a grounded projection.

Input Data:

- **Cumulative Thoughts**: {(thought_history)}
- **Current Speech**: {(Talking_block)}

Output Format: Respond strictly with a JSON object:

```
{
"alignment_analysis": "Analysis of whether speech content ⊆ thought content...",
"hallucination_detected": true/false,
"consistency_score": float (0.0 for ungrounded, 1.0 for fully anchored)
}
```

Figure 10: Prompt specification for the Internal-External Consistency Reward (R_{const}). This prompt enforces "Logical Anchoring" by verifying that the spoken content is a subset of the cumulative thought history, strictly penalizing hallucinated increments where speech outpaces reasoning.

Chain-of-Thought Evaluation Prompt for Sparse Logic Flow Reward

You are a **Global Coherence Auditor** and **Long-Range Logic Validator**. Your purpose is to enforce strict global consistency across non-adjacent generation steps.

Goal: Execute "Cross-Step Logical Coherence Check" by analyzing a randomly sampled pair of distal blocks (B_i, B_j) to detect long-range semantic drift.

Evaluation Criteria (Global Consistency Verification):

1. **Antecedent-Consequent Fidelity**
   - **Goal**: Verify reasoning trajectory remains faithful to early constraints.
   - **Execution**:
   - Compare premises established in B_i with the trajectory in B_j.
   - Ensure the later block does not retcon or alter fundamental axioms.
   - Detect if distal content ignores state variables set in the antecedent.
2. **State Space Integrity (Variable Consistency)**
   - **Goal**: Scrutinize specific entities and numerical values.
   - **Execution**:
   - Match variable assignments in B_j against the exact definitions in B_i.
   - Identify "silent mutations" where values change without operation.
   - Flag contradictions in numerical constants or entity attributes.

Task: Compare the two non-adjacent blocks and output a flow score.

Input Data:

- **Block i** (Time $t - \Delta$): {(block_i_content)}
- **Block j** (Time t): {(block_j_content)}

Output Format: Respond strictly with a JSON object:

```
{
"coherence_analysis": "Analysis of logical continuity...",
"conflict_detected": true/false,
"flow_score": float (0.0 for contradiction, 1.0 for consistency)
}
```

Figure 11: Prompt specification for the Sparse Logic Flow Reward (R_{flow}). This prompt implements the "Sparse Temporal Sampling" mechanism, directing the evaluator to audit non-adjacent block pairs for global variable consistency and logical adherence to antecedent premises.

Chain-of-Thought Evaluation Prompt for Multimodal Outcome Reward

You are a **Dual-Domain Quality Arbiter** specializing in both **Mathematical/Logical Verification** and **Acoustic-Affective Analysis**.

Goal: Assess the final output quality by strictly balancing "High-IQ" (Logical Correctness) and "High-EQ" (Acoustic Expressiveness).

Input Data:

- **Context**: Question Text: {(q_text)} | Question Audio
- **Response**: Answer Text: {(a_text)} | Answer Audio
- **Reference**: Ground Truth: {(ground_truth)}

Evaluation Criteria (Dual-Stream Assessment):

1. **Logic-Truth Alignment (Textual Accuracy)**

- **Goal**: Verify objective correctness against Ground Truth.
- **Execution**:
- Rigorously compare the Answer Text with the Ground Truth.
- Identify if the reasoning path is valid and leads to the correct solution.
- Penalize factual errors or calculation mistakes, regardless of vocal confidence.

2. **Acoustic-Scenario Fit (Expressiveness)**

- **Goal**: Evaluate if Answer Audio tonally matches the context.
- **Execution**:
- Analyze prosody (pitch, speed, pause) relative to the detected emotional state.
- Reward "Adaptive Resonance": e.g., soothing tone for anxious queries.
- Penalize "Mechanical Flatness" or robotic delivery lacking fluctuation.
- Penalize "Tonal Dissonance": e.g., mismatched cheerful tone for serious errors.

Task: Synthesize both streams and output a weighted outcome score.

Output Format: Respond strictly with a JSON object:

```
{
"logic_verification": "Analysis of correctness vs ground truth...",
"acoustic_analysis": "Analysis of tonal fit to context...",
"outcome_score": float (0.0 to 1.0)
}
```

Figure 12: Prompt specification for the Multimodal Outcome Reward ($R_{outcome}$). This prompt instructs the evaluator to function as a dual-domain expert, verifying factual accuracy against ground truth while simultaneously assessing the acoustic expressiveness and emotional appropriateness of the synthesized speech.

User Query:
Consider a cubic polynomial function, defined as the cube of the independent variable, minus three times the variable, plus the constant two. First, determine the precise coordinates of all points where the instantaneous slope is zero. Second, identify which of these points represents a local maximum. Finally, calculate the area of the enclosed region bounded by the function's graph, the horizontal axis, and vertical lines extending downwards from the origin and the local maximum's x-coordinate respectively.

<THINK>

1. **<ROLE_AFFECTIVE_STRATEGY>**
   - **Goal**: Pre-generate the emotional tone to stabilize prosody.
   - **Execution**:
     - Role Identity: A patient, methodical senior mathematics professor.
     - Emotional Tone: Natural, calm, authoritative yet encouraging.
     - Linguistic Style: Moderate speed, clear articulation with deliberate pauses.
2. **<PROBLEM_DECONSTRUCTION>**
   - **Goal**: Establish a clear symbolic space.
   - **Execution**:
     - Extract entities: Function $f(x) = x^3 - 3x + 2$; constraints $f'(x) = 0$; local max identification; definite integral $\int f(x)dx$.
     - Input/Output: Natural language problem $\rightarrow$ Natural language answer with precise values.
     - Transform: Symbolize $\rightarrow$ Differentiate $\rightarrow$ Solve Critical Points $\rightarrow$ Integrate.
3. **<TASK_POSITIONING>**
   - **Goal**: Assess complexity and plan solution path.
   - **Execution**:
     - Cognitive Load: High (Multi-step calculus).
     - Computational Budget: 8 logical steps required.
     - Path: Define $f(x)$ $\rightarrow$ Find $f'(x)$ $\rightarrow$ Solve $f'(x) = 0$ $\rightarrow$ Evaluate points $\rightarrow$ Find Max $\rightarrow$ Setup Integral $\rightarrow$ Compute Area.
4. **<SYMBOLIC_REASONING>**
   - **Goal**: Execute deduction and verification.
   - **Execution**:
     - Step 1: Define $f(x) = x^3 - 3x + 2$.
     - Step 2: Differentiate. $f'(x) = 3x^2 - 3$.
     - Step 3: Solve $3x^2 - 3 = 0 \Rightarrow 3(x^2 - 1) = 0 \Rightarrow x = 1, x = -1$.
     - Step 4: Evaluate. $f(1) = 1 - 3 + 2 = 0 \Rightarrow (1, 0)$. $f(-1) = -1 + 3 + 2 = 4 \Rightarrow (-1, 4)$.
     - Step 5: Identify Max. $4 > 0$, so Local Max is at $(-1, 4)$. Bounds: $[-1, 0]$.
     - Step 6: Antiderivative $F(x) = \frac{x^4}{4} - \frac{3x^2}{2} + 2x$.
     - Step 7: Area $A = F(0) - F(-1) = 0 - (\frac{1}{4} - \frac{3}{2} - 2) = -(\frac{1-6-8}{4}) = \frac{13}{4} = 3.25$.

</THINK>

Response:
The points where the function's slope is zero are minus one comma four, and one comma zero. The local maximum occurs at the point minus one comma four. The calculated area of the enclosed region is thirteen-fourths, or precisely three point two five.

Figure 13: Case study of the model processing a complex calculus problem using the R-P-T-S paradigm. The <THINK> block demonstrates the transition from affective strategy setting to symbolic execution, ensuring prosodic stability and logical accuracy.