DEIRD: A DECOUPLED CONTINUOUS IMPLICIT REASONING FRAMEWORK FOR DIALOGUE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

By leveraging the "think before speaking" response paradigm, large models can generate a Chain-of-Thought (CoT) prior to answering to enhance their deep reasoning capabilities. However, directly applying this paradigm to spoken dialogues poses three severe challenges: huge training costs caused by the modality gap, the unacceptable latency introduced by the "think before speaking" mechanism, and limited reasoning exploration depth. To address these issues, we propose **DeIRD**, a latency-controllable and low-training-cost implicit reasoning framework for spoken dialogue models that simultaneously achieves "think and speak decoupled" and "think before speaking". This framework is primarily realized through three core components: 1) We train a text VAE to compress variable-length textual CoT sequences into fixed-length continuous latents, serving as the carrier for implicit reasoning; 2) We introduce a continuous masked prediction model **DMIR** that takes user speech as input and outputs a fixed-length sequence of these latent targets via iterative confidence-guided "mask-predict" decoding, to achieve "think and speak decoupled" and "think before speaking"; 3) We construct an implicit reasoning training dataset, **LatentReason-40k**, and find that the downstream backbones just require one epoch of end-to-end fine-tuning to acquire superior reasoning capabilities compared to explicit reasoning models trained on the same volume of data. Compared with various explicit-CoT baselines, DeIRD notably reduces the Time to First Audio (TTFA) while improving general dialogue quality, style control, deep reasoning, and reasoning diversity.

1 INTRODUCTION

Large language models improve performance on complex tasks by generating explicit text as intermediate reasoning steps, adopting a "think before speaking" approach (Kojima et al., 2022; Wang et al., 2023; Wei et al., 2022). However, endowing spoken dialogue models (Ji et al., 2024; Xu et al., 2025; Team et al., 2025) with the "think before speaking" capability faces three major limitations. First, we hypothesize that the modality gap between speech and text may limit the effectiveness of prompting spoken dialogue models to generate reasoning before answering, potentially increasing reliance on substantial amounts of CoT training data to acquire explicit reasoning capabilities beyond those elicited through prompting alone in speech interactions. Second, given the real time requirements of voice interaction, generating complete explicit and discrete CoT sequences brings unacceptable latency. Finally, explicit CoT performs deterministic sampling at high entropy positions, which limits the exploration of the model into potential reasoning paths (Xu et al., 2026; Du et al., 2026a). Compared to "think before speaking", recent studies have attempted to use "think and speak decoupled" dual brain architectures or "think while speaking" or "think while listening" mechanisms to mitigate TTFA latency (Chiang et al., 2025b;a), but two drawbacks still remain: 1) most of them still rely on *discrete* and *variable-length* explicit CoT tokens, which restricts the reasoning speed and depth of the models; 2) they all require "thought chunking" to achieve streaming interaction between the front and back brains or the "speak while thinking" method, which results in the early outputs of the model lacking sufficient thought, thereby affecting the response quality.

To overcome these issues, spoken dialogue models urgently need a new implicit thinking framework: under the premise of ensuring controllable latency, it can achieve the complete thinking of "think

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

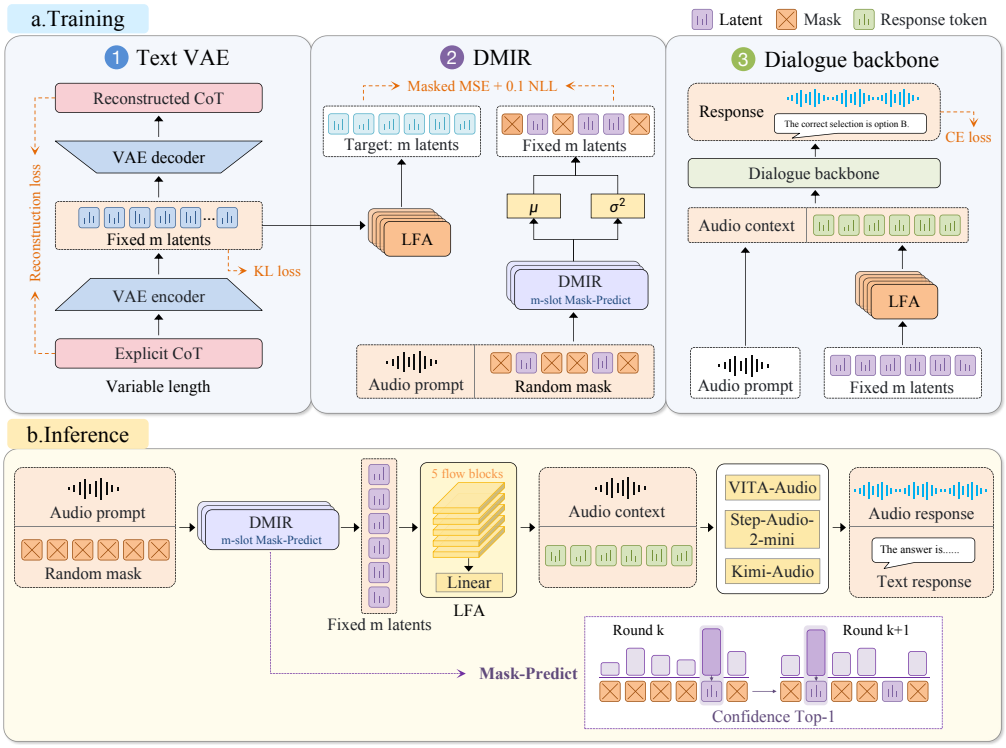


Figure 1: **Architecture of the DeIRD framework. (a) Training**: 1) A text VAE encoder compresses variable-length explicit CoT into m fixed latents, optimized via reconstruction and KL losses. 2) LFA aligns these latents as targets for DMIR, which learns the audio-to-latent mapping via Masked MSE and NLL losses. 3) The dialogue backbone is fine-tuned via CE loss to generate text and audio responses conditioned on these latents. **(b) Inference**: Given an audio prompt, DMIR iteratively generates m implicit latents via a confidence-guided Top-1 mask-predict mechanism. LFA then adapts these latents to guide downstream backbones in response generation.

before speaking", and only requires a small training cost to help the spoken dialogue base model acquire reasoning capabilities. Therefore, we propose the **DeIRD** (**De**coupled **I**mplicit **R**easoning for **D**ialogue): an implicit reasoning framework that simultaneously achieves "think and speak decoupled" and "think before speaking", with controllable latency and low training cost. This framework uses thinking latent variables as thinking content, and decouples the reasoning capability from the backbone dialogue model via a separately continuous masked prediction model, **DMIR** (**D**ialogue conditioned **M**asked **I**mplicit **R**easoner), to achieve the separation of thinking and answering (Zhou et al., 2025; 2026). DMIR receives audio input and outputs implicit thinking latent variables with a fixed length, which together with the audio prompt guide the end-to-end training of the dialogue model. The downstream dialogue model only requires 1 epoch of end-to-end training on the LatentReason 40k dataset to achieve better reasoning capabilities than explicit thinking trained on the same amount of data. Figure 1 b shows the synergistic interaction mechanism between the DMIR implicit reasoning module and the dialogue model backbone: in the thinking phase, DMIR generates a fixed number of implicit latent variables based on the input audio; in the answering phase, to match hidden-space shapes and dimensions across models, we use **LFA** (**L**atent **F**low **A**dapter) (Tu et al., 2026) to transform this latent variable sequence before feeding it into the dialogue model together with the audio prompt. **This design has two core advantages.** First, based on "think and speak decoupled", the independent DMIR generates implicit reasoning content, achieving "think before speaking" and solving the problem of insufficient thinking in the early responses of existing methods. Second, DMIR adopts a masked prediction paradigm, generating a *fixed length* implicit latent variable sequence in a continuous space based on a preset number of mask slots. This paradigm not only ensures controllable reasoning latency, but also avoids the drawback of discrete tokens prematurely discarding other potential reasoning paths at reasoning branch points.

During the training phase, we train a text VAE and an implicit reasoning model DMIR. The text VAE is responsible for compressing variable length explicit text sequences into a fixed number of latent distributions (Liu & Qiu, 2025); these high density latent variables, after being adapted to match hidden-space shapes and dimensions via the LFA module, are mapped to the original feature space of DMIR to provide supervision targets for DMIR. DMIR then learns the mapping from the audio question to these latent targets. In summary, the main contributions of this paper are as follows:

- We propose DeIRD, a latency-controllable, low-training-cost implicit reasoning framework for spoken dialogue models that simultaneously achieves "think and speak decoupled" and "think before speaking", and construct the accompanying **LatentReason-40k training dataset**. During downstream training, this framework does not require additional loss to supervise the thinking process and only needs 1 epoch of training, featuring broad adaptability and low training cost; during inference, it ensures the reasoning depth of the model and reduces interaction latency.
- We train a text VAE to compress and reconstruct variable length explicit text sequences through fixed length latent variables, providing target latent sequences for downstream models. We also introduce DMIR, a continuous masked prediction model that outputs fixed length implicit latent variables based on the input audio, achieving the separation of thinking and answering.
- We evaluate DeIRD on three spoken dialogue backbones covering interleaved single-stream and parallel dual-head architectures. Results show improved dialogue performance, style control, and reasoning performance, with lower TTFA than the evaluated explicit CoT baselines.

2 Related Work

Current research on spoken dialogue reasoning focuses on two directions: introducing thinking capabilities and reducing latency. **To introduce thinking capabilities,** early studies construct explicit text thoughts based on user audio semantics, using SFT to train models to output step-by-step analyses before answering (Ma et al., 2025; Xie et al., 2025). Recent studies optimize output preferences via Reinforcement Learning (RL), focusing thoughts more on acoustic features alongside semantics (Rong et al., 2025; Tian et al., 2025). However, this explicit "think before speaking" paradigm must completely output thoughts before answering, violating the low-latency requirements of voice interaction. **To reduce thinking latency,** Thinking Machines LAB [1] and MPS (Wu et al., 2026a) proposed a "think and speak decoupled" dual-brain architecture to reduce TTFA latency; Stitch (Chiang et al., 2025b), Shanks (Chiang et al., 2025a), InterRS (Du et al., 2026b), FLAIR (Wu et al., 2026b) and Dual-reasoner (Li et al., 2026) explored streaming mechanisms like "think while speaking" or "think while listening," interweaving thoughts into audio reception or generation. Although these methods effectively reduce latency, they all rely on "thought chunking": the former uses thought chunking to achieve interaction between the front and back brains, while the latter requires inserting thought chunks during the answering phase. This design causes the early responses of the model to lack sufficient thinking, compromising the quality of the responses. Meanwhile, these methods alter original output paradigms, forcing models to rely on large-scale data and longer training to adapt. **In contrast,** DeIRD is a latency-controllable, low-training-cost implicit reasoning framework that simultaneously achieves "think and speak decoupled" and "think before speaking".

3 Method

3.1 Overview

Figure 1 illustrates the overall workflow of the DeIRD training framework. The top section depicts the training pipeline, which includes the training of the text VAE, the DMIR model, and the spoken dialogue backbones of various architectures. The bottom section outlines the inference pipeline, demonstrating how dialogue models of varying architectures generate responses under the guidance of the implicit reasoning latents provided by DMIR.

Current spoken dialogue models predominantly employ two speech and text generation paradigms: **interleaving,** which generates a single interleaved token stream, and **parallel,** which generates text and speech streams with coupled states. For consistency, we unify the mixed-modal output of the

[1] https://thinkingmachines.ai/blog/interaction-models/

spoken dialogue models as a sequence of decoding steps $R_{1:t}$. Traditional explicit Chain-of-Thought (CoT) frameworks rely on explicit discrete text tokens and consist of two stages: the reasoning stage and the answering stage. In the reasoning stage, given the user audio input p, the model generates a discrete text sequence $d_{1:L}$ of length L as the problem analysis. In the answering stage, conditioned on the user input p and the intermediate analysis sequence $d_{1:L}$, the model computes the joint probability to generate the final response $R_{1:t}$:

$$P_\theta(R_{1:t} \mid p, d_{1:L}) = \prod_{i=1}^{t} P_\theta(R_i \mid p, d_{1:L}, R_{1:i-1}) \tag{1}$$

However, the generation of explicit CoT sequences relies on deterministic sampling, which not only causes the model to lose potential reasoning paths but also fails to achieve controllable reasoning length. Therefore, the DeIRD framework uses fixed-length, continuous-space implicit latent sequences to replace explicit CoT sequences for carrying reasoning information. This framework consists of three stages: First, training the text VAE. We train a text VAE designed for fixed-length compression. Its encoder compresses a variable-length discrete text sequence $d_{1:L}$ ($L < 2048$) into a latent distribution $z_{1:m}$ of length m. During inference, the decoder auto-regressively reconstructs the original text sequence relying exclusively on this latent distribution $z_{1:m}$. Second, training the DMIR model. Based on DIFFA-2, we train a speech comprehension model, DMIR. Using the mask-predict mechanism, this model learns the mapping from the user audio input p to the corresponding VAE-compressed latents $z_{1:m}$. Throughout this process, the predefined number of mask slots m in DMIR is kept consistent with the compression length of the VAE latents. Finally, training the Spoken Dialogue Backbone. Using the LatentReason-40k dataset, the implicit reasoning latent sequence is adapted to match hidden-space shapes and dimensions via the LFA module, and is then inserted into the input context. Fine-tuning the spoken dialogue backbone end-to-end for only a single epoch is sufficient to help the base model acquire implicit reasoning capabilities.

3.2 WORKFLOW OF DEIRD

3.2.1 TEXT VAE

The text VAE compresses variable-length ($L < 2048$) explicit discrete Chain-of-Thought (CoT) text sequences $d_{1:L} = (d_1, \ldots, d_L)$ into m fixed-length continuous latent variables $z_{1:m}$. Built upon the deterministic C3 architecture, the Encoder LM appends m learnable query embeddings at the end of the input sequence as placeholders. By extracting the final hidden states at these m positions, it computes the diagonal Gaussian posterior distribution as follows:

$$q_\phi(z_{1:m} \mid d_{1:L}) = \mathcal{N}(z_{1:m}; \mu, \sigma^2) \tag{2}$$

During the reconstruction phase, relying exclusively on the latent sequence $z_{1:m}$, the Decoder LM auto-regressively maximizes the conditional probability of reconstructing the original text sequence:

$$p_\theta(d_{1:L} \mid z_{1:m}) = \prod_{i=1}^{L} p_\theta(d_i \mid z_{1:m}, d_{1:i-1}) \tag{3}$$

3.2.2 DMIR

The objective of DMIR is to learn the mapping from the user audio input p to implicit reasoning latents. We initialize DMIR from DIFFA-2's backbone and replace its vocabulary prediction head with a Gaussian distribution prediction head. DMIR is an audio-conditioned masked predictive model that operates independently of the spoken dialogue backbone. Its Mask-Predict mechanism achieves fixed-length generation through a predefined number of mask slots, thereby ensuring controllable reasoning latency. During inference, given the input audio features p, the model initializes from m fully masked slots. In each iteration, conditioned on the audio context p and the current visible slot features z_{vis}, the model predicts the Gaussian means and log-standard deviations for all masked slots in parallel, and computes the confidence score c_k:

$$(\hat{\mu}_{1:m}, \log \hat{\sigma}_{1:m}) = g_\theta(p, z_{\text{vis}}, M), \quad c_k = -\frac{1}{d} \sum_{j=1}^{d} \log \hat{\sigma}_{k,j} \tag{4}$$

where M denotes the mask status. The model utilizes the negative mean of the predicted log-standard deviations as the confidence metric, retaining the uncommitted slot with the highest confidence (Top-1) in each round and committing its mean as the continuous feedback condition for the

subsequent round. After m rounds of confidence-guided iterative one-slot unmasking, the model outputs a latent sequence of length m, which serves as the final implicit reasoning context. **Furthermore,** fully connected layers project dimensionally distinct VAE compressed latents into the DMIR token space for masking. Strictly matching and aligning the number of latent slots effectively prevents sequential correspondence errors during feature mapping.

3.2.3 SPOKEN DIALOGUE BACKBONE

The spoken dialogue backbone retains its native generation paradigm, treating the implicit latent sequence $\hat{z}_{1:m}$ output by DMIR as an additional soft prompt. Conditioned on the user audio input p and this latent sequence $\hat{z}_{1:m}$, the model computes the probability of generating the joint response $R_{1:t}$, which comprises the text sequence $T_{1:b}$ and the audio sequence $A_{1:a}$:

$$P_\theta(R_{1:t} \mid p, \hat{z}_{1:m}) = \prod_{i=1}^{t} P_\theta(R_i \mid p, \hat{z}_{1:m}, R_{1:i-1}) \tag{5}$$

The advantage of this design is that, although various spoken dialogue models possess different generation paradigms, they can uniformly reuse the implicit reasoning features generated by DMIR and aligned via the LFA, without requiring modifications to their native architectures or the introduction of auxiliary reasoning loss functions to supervise the thought process.

3.2.4 LATENT FLOW ADAPTER

The primary objective of the Latent Flow Adapter (LFA), adapted from Tu et al. (2026) to our spoken dialogue framework, is to match hidden-space shapes and dimensions across models, participating in the gradient backpropagation and parameter updates of the backbone model during training. In the DeIRD framework, the LFA is independently instantiated and deployed at two locations: first, within DMIR, to establish the mapping between the original continuous VAE latent space and the Gaussianized mask prediction space; second, at the input of the downstream spoken dialogue model, to adapt the predicted implicit latents to the hidden feature space of specific backbone models.

Architecturally, the LFA is composed of a 5-block slot-wise affine-coupling flow in series with a linear projector. To match hidden-space shapes, the affine-coupling flow F maps the original latent variable e to the Gaussianized feature u via a reversible transformation $u = F(e)$. The probability density transformation within this continuous space satisfies the following relationship:

$$\log p_E(e) = \log p_U(F(e)) + \log |\det J_F(e)| \tag{6}$$

This lossless transformation $e = F^{-1}(u)$ applies to the 5-block affine-coupling flow, transforming latent features to match hidden-space shapes before a non-injective linear maps these features to the hidden dimension, thereby adapting to the architectural requirements of different backbones.

3.3 TRAINING

3.3.1 TEXT VAE TRAINING

The text VAE is initialized with the pre-trained C3 VAE. The training objective of this model is to minimize the negative Evidence Lower Bound (ELBO) of a β-VAE, augmented with sequence perturbation regularization. The total loss function is formulated as follows:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q_\phi(z_{1:m}|\tilde{d}_{1:L})}\big[\log p_\theta(d_{1:L} \mid z_{1:m})\big] + \beta D_{\text{KL}}\big(q_\phi(z_{1:m} \mid \tilde{d}_{1:L}) \,\|\, p_{\text{base}}(z_{1:m})\big) \tag{7}$$

where $\tilde{d}_{1:L}$ represents the input text sequence subjected to random masking at a ratio of 15% on the encoder side. Because the decoder employs an auto-regressive architecture, the training process relies on teacher-forcing. This random masking strategy is designed to mitigate the posterior collapse phenomenon induced by auto-regressive decoding, thereby compelling the model to reconstruct the complete original text relying exclusively on the latent variables $z_{1:m}$ during inference.

3.3.2 DMIR TRAINING

DMIR is initialized with the pre-trained DIFFA-2. We replace the discrete vocabulary mapping head with a continuous Gaussian prediction head. The supervision ground truth, $u = F(\mu)$, is

the Gaussianized feature transformed via the LFA module from the posterior mean μ output by the text VAE. During training, we sample a random mask ratio $t \sim U(0, 1)$ for each instance and independently perform Bernoulli mask sampling for the m slots accordingly, yielding the mask status $M_k \in \{0, 1\}$ for each slot. Subsequently, conditioned on the user audio input p and the unmasked ($M_k = 0$) continuous feedback slots, the model predicts the Gaussian means $\hat{\mu}$ and log-standard deviations $\log \hat{\sigma}$ for all slots. The loss function is computed exclusively on the set of masked slots $\mathcal{M} = \{k \mid M_k = 1\}$. The overall optimization objective $\mathcal{L}_{\text{DMIR}}$ combines the Mean Squared Error (MSE) and the Gaussian Negative Log-Likelihood (NLL) for uncertainty estimation:

$$\mathcal{L}_{\text{DMIR}} = \mathcal{L}_{\mu} + 0.1\mathcal{L}_{\text{NLL}} \tag{8}$$

$$\mathcal{L}_{\mu} = \frac{1}{|\mathcal{M}| \cdot d} \sum_{k \in \mathcal{M}} \sum_{j=1}^{d} (u_{k,j} - \hat{\mu}_{k,j})^2 \tag{9}$$

$$\mathcal{L}_{\text{NLL}} = \frac{1}{|\mathcal{M}| \cdot d} \sum_{k \in \mathcal{M}} \sum_{j=1}^{d} \left[\frac{1}{2} \left(\frac{u_{k,j} - \text{sg}(\hat{\mu}_{k,j})}{\hat{\sigma}_{k,j}} \right)^2 + \log \hat{\sigma}_{k,j} \right] \tag{10}$$

where d denotes the single-slot channel dimension of the latents, and $\text{sg}(\cdot)$ represents the stop-gradient operation. This operation is designed to prevent the model from evading mean regression penalties by anomalously amplifying the standard deviation $\hat{\sigma}_{k,j}$, compelling the predicted scale to fit the mean residuals, which acts as a confidence proxy for slot selection during inference.

3.4 Spoken Dialogue Backbone Training

DeIRD introduces implicit reasoning to the dialogue backbone without additional supervision losses. The conditional latent sequence pre-generated by DMIR, after being aligned via the LFA module, is concatenated into the input context as a soft prompt. Detailed insertion mechanisms are provided in Appendix A. Conditioned on the input audio p and the aligned latent sequence $\hat{z}_{1:m}$, the backbone model predicts the mixed-modal target response $R_{1:t}$, achieving end-to-end training solely relying on the CE loss. Based on the backbone's generative architecture, the loss is formulated as:

$$\mathcal{L}_{\text{interleaved}} = -\sum_{i=1}^{t} \log P_{\theta}(R_i \mid p, \hat{z}_{1:m}, R_{1:i-1}) \tag{11}$$

$$\mathcal{L}_{\text{parallel}} = -\sum_{i=1}^{t} [\log P_{\theta}(T_i \mid p, \hat{z}_{1:m}, T_{1:i-1}) + \log P_{\theta}(A_i \mid p, \hat{z}_{1:m}, A_{1:i-1}, \mathcal{C})] \tag{12}$$

where $\mathcal{C}$ denotes the cross-modal context relied upon for audio generation in the parallel architecture. During training, the parameters of the LFA module are updated synchronously. The placeholder slots occupied by the latent variables $\hat{z}_{1:m}$ do not calculate any loss themselves; all supervision signals are derived directly from the target text and audio mixed sequence $R_{1:t}$.

3.5 LatentReason-40k

To endow the spoken dialogue backbone with implicit reasoning capabilities at an extremely low training cost of a single end-to-end epoch, we construct LatentReason-40k, a high-quality dataset comprising 46,809 samples (685 hours) across two subsets: Semantic Content and Paralinguistic.

Data Composition and Filtering. The Semantic Content subset encompasses 8 categories: Math, Science QA, Instruction, Logic, Code, Creative, Ultra, and Dialogue. Samples undergo a four-step filtering pipeline: (1) rule-based filtering to remove length violators and speech-incompatible visual layouts; (2) LLM-based domain classification to eliminate visually-dependent samples unsolvable solely via audio; (3) employing Qwen2.5-Omni (Xu et al., 2025) to retain moderate-difficulty samples with empirical accuracy in $[0.125, 0.875]$; (4) synthesizing filtered text into high-fidelity audio via TTS. Derived from WavReward positive samples, the Paralinguistic subset includes 3 categories (explicit understanding, explicit generation, and implicit) covering 8 acoustic features: Emotion, Gender, Pitch, Volume, Age, Accent, Speed, and Language. Details appear in Appendix B.

Explicit Chain-of-Thought Construction. We utilize Gemini 3.1 Pro[2] to generate explicit Chain-of-Thought (CoT) sequences for all samples. To ensure high-quality fixed-length text VAE compression, CoT length is limited to under 1,500 words. All CoTs follow a unified template comprising

[2]https://deepmind.google/models/gemini/pro/

Table 1: **Downstream Policy Evaluation and Ablation Studies.** We compare the DeIRD framework against multiple baselines and validate its design choices across four ablation settings. In the columns, **S**, **V**, and **K** denote the backbone models: Step-Audio-2-mini, VITA-Audio, and Kimi-Audio-7B-Instruction. In the main results, best results are **bolded** and second best are underlined.

Method	VoiceBench						VStyle			GSM8K			OpenAudioBench					
	IFEv. ↑			Avg ↑			Avg ↑			P@8 ↑			Avg ↑			TTFA ↓		
	S	V	K	S	V	K	S	V	K	S	V	K	S	V	K	S	V	K
Main Results: Downstream Policy Evaluation																		
Base Model	42.2	27.7	61.5	58.7	50.8	72.8	3.08	2.52	3.06	71.3	88.9	96.1	55.1	52.4	70.2	**0.37**	**0.11**	**0.22**
+ DIFFA2 CoT (Free)	26.0	22.8	37.8	40.0	44.7	50.8	1.12	1.71	2.19	54.6	60.8	84.0	34.4	32.8	54.9	14.2	12.8	16.3
+ DIFFA2 CoT	54.4	34.5	63.8	62.4	53.0	70.8	2.44	2.75	2.87	70.2	78.9	94.7	58.2	49.9	65.6	14.2	12.8	16.3
+ Prompt (Free)	35.7	30.4	55.2	60.3	58.9	71.9	3.08	2.58	2.88	68.7	88.8	96.1	61.5	56.1	69.5	0.42	0.13	0.26
+ Explicit CoT	45.5	39.6	64.1	62.5	58.7	73.0	2.54	2.81	3.14	78.4	85.4	95.9	55.7	55.9	67.6	18.6	9.13	12.23
+ DeIRD (Ours)	**57.4**	**56.0**	**70.0**	**64.9**	**60.7**	**74.4**	**3.15**	**2.91**	**3.18**	**88.1**	**91.0**	**97.2**	**62.6**	**57.6**	**70.8**	1.50	1.57	1.67
Ablation 1: Semantic Validity of Latent Variables																		
Base Model	42.2	27.7	61.5	58.7	50.8	72.8	3.08	2.52	3.06	71.3	88.9	96.1	55.1	52.4	70.2	0.37	0.11	0.22
+ Noise Latent	44.7	55.6	68.0	60.7	59.6	71.7	2.84	2.78	1.63	76.6	88.1	94.8	57.7	56.3	68.0	0.37	0.12	0.22
Ablation 2: Optimal Parameter Exploration																		
+ DeIRD (Ours)	57.4	56.0	70.0	64.9	60.7	74.4	3.15	2.91	3.18	88.1	91.0	97.2	62.6	57.6	70.8	1.50	1.57	1.67
+ DeIRD (2048, 64)	55.8	55.8	66.6	63.8	59.9	71.5	2.85	2.82	1.72	87.3	88.0	95.3	62.4	55.1	67.2	2.98	3.03	3.16
+ DeIRD (2048, 16)	58.0	55.5	69.0	63.3	59.7	72.4	2.90	2.81	3.14	88.1	87.9	95.4	62.5	55.4	69.5	0.80	0.86	0.96
+ DeIRD (3584, 32)	58.1	56.5	67.8	64.1	59.3	74.2	2.73	2.84	3.22	87.0	88.4	95.1	61.6	55.4	69.4	1.51	1.57	1.67
+ DeIRD (768, 32)	54.2	56.0	67.6	63.7	60.1	71.7	2.87	2.66	3.29	86.1	88.0	95.5	62.5	56.2	68.6	1.50	1.56	1.66
Ablation 3: w/o LFA Module																		
+ DeIRD (Ours)	57.4	56.0	70.0	64.9	60.7	74.4	3.15	2.91	3.18	88.1	91.0	97.2	62.6	57.6	70.8	1.50	1.57	1.67
- LFA Module	56.6	53.5	66.8	63.8	60.0	72.3	2.81	2.84	1.54	86.8	88.6	95.2	60.3	56.6	67.2	1.50	1.55	1.66
Ablation 4: Decoding Strategy Comparison																		
Top 1 Decoding (Ours)	57.4	56.0	70.0	64.9	60.7	74.4	3.15	2.91	3.18	88.1	91.0	97.2	62.6	57.6	70.8	1.50	1.57	1.67
Top 2 Decoding	57.0	54.0	68.1	64.3	59.8	72.6	2.84	2.78	1.61	87.0	87.4	95.6	62.4	56.6	68.0	0.81	0.87	0.97
Top 4 Decoding	56.5	57.4	69.2	64.5	60.4	72.4	2.88	2.83	1.61	88.3	88.1	95.0	62.3	56.1	63.7	0.47	0.52	0.62

three components: an analysis of input audio's paralinguistic features *(e.g., "the user spoke in an angry tone")*, a logical deduction of the question's semantic content, and acoustic paralinguistic specifications for the final response *(e.g., "a gentle tone is required for the response")*.

4 Experiments

4.1 Experimental Setup

Datasets. DeIRD training involves three stages. First, the text VAE trains on 401,019 mixed samples combining SYNTHETIC-2 CoT data under 4,096 tokens and LatentReason-40k explicit CoT trajectories. Second, DMIR trains on LatentReason-40k, utilizing question audio as input and the fixed-length latent sequences offline-compressed by the VAE as supervision targets. Finally, the spoken dialogue backbones undergo end-to-end supervised fine-tuning directly on LatentReason-40k.

Baselines. To validate the effectiveness and adaptability of the DeIRD, we select three dialogue models covering both generation paradigms as backbones. These comprise Step-Audio-2-mini (Wu et al., 2025) and VITA-Audio-Plus-Vanilla (Long et al., 2025), using single-stream interleaved generation, and Kimi-Audio-7B-Instruct (KimiTeam et al., 2025), which adopts a parallel design.

Evaluation Metrics. We evaluate the model's comprehensive capabilities using four benchmarks and one latency metric. **VoiceBench** (Chen et al., 2025) examines the model's general dialogue capabilities. **VStyle** (Zhan et al., 2025) evaluates speech style control. For reasoning, **OpenAudioBench** (Li et al., 2025b) tests complex logical reasoning to validate gains from the implicit thinking phase; **GSM8K** (Cobbe et al., 2021) measures reasoning depth and diversity via pass@k ($k \in \{1, 2, 4, 8\}$) accuracy, sampled at temperature 0.7. Since implicit reasoning prevents explicit discrete sampling from prematurely truncating potential paths, the pass@k metric based on multiple independent executions reflects the model's advantage in exploring the solution space. We calculate the **Time to First Audio (TTFA)** on OpenAudioBench, a metric measuring the time from the end of user input to the output of the first audio token, which reflects the model's actual inference speed.

Table 2: VAE reconstruction performance and DMIR query relevance across varying configurations.

Dimension (d)	Slots (m)	Text VAE Recon.		DMIR Interp.
		Token ACC (↑)	WER (↓)	R_{rel} (↑)
2048 (DeIRD)	**32**	99.9%	0.003	99.5%
2048	64	99.9%	0.023	98.5%
2048	16	99.7%	0.048	98.5%
3584	32	99.9%	0.003	99.0%
768	32	95.4%	0.266	94.0%

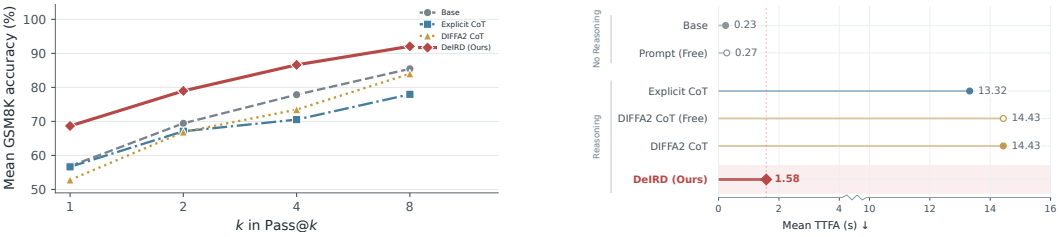


Figure 2: **Left:** Mean GSM8K across the three models; **Right:** Mean TTFA across the three models.

4.2 Evaluation Protocol

Reconstruction and Query Relevance. We set five parameter configurations for latent dimension d and quantity m: $(d, m) \in \{(2048, 32), (2048, 64), (2048, 16), (3584, 32), (768, 32)\}$, and the primary model adopts $(2048, 32)$. We evaluate through two steps: 1) using WER and reconstruction accuracy (ACC) to evaluate the reconstruction effect of the text VAE; 2) utilizing Gemini 3.1 Pro to evaluate the thinking text decoded from the latent space, to verify the query relevance of the latent variables predicted by DMIR. Let N be the total number of queries and $I_n \in \{0, 1\}$ indicate whether the decoded text of the nth sample is relevant to the query. We define the query relevance score R_{rel}:

$$R_{\text{rel}} = \frac{1}{N} \sum_{n=1}^{N} I_n. \tag{13}$$

Downstream Policy Evaluation. We compare the DeIRD framework with four baseline methods across three downstream dialogue base models: 1) **Base Model**: the original non reasoning version of the backbone models; 2) **DIFFA2 CoT (Free) / DIFFA2 CoT**: a contrastive group inserting explicit CoT generated by DIFFA-2, which is evaluated under both training free and post training fine tuning settings; 3) **Prompt (Free)**: a zero shot prompting baseline, which solely requires the model to auto regressively generate explicit CoT via prompt instructions; 4) **Explicit CoT**: an explicit fine tuning baseline that learns to directly output the text CoT from the training data before generating the response. We report the performance of the above methods on four benchmarks and the TTFA on OpenAudioBench, to prove the DeIRD implicit reasoning framework is superior to existing speech reasoning schemes in both comprehensive performance and inference speed. All fine-tuned methods use the same backbone, training data, and one epoch of training within each backbone comparison.

Ablation Studies. To verify that the performance gain of DeIRD originates from the implicit reasoning framework that simultaneously achieves "think and speak decoupled" and "think before speaking", we designed four sets of experiments: 1) Semantic validity of latent variables: replacing the input latent variables with $(2048, 32)$ noise close to the expected space of the vocabulary for training and inference, to exclude the illusion of performance improvement caused by the delayed response time brought by continuous latents; 2) Optimal parameter exploration: respectively training the VAE and DMIR models under the aforementioned five (d, m) parameters, and reporting their downstream performance on VoiceBench, OpenAudioBench, and VStyle; 3) Necessity of the LFA module: removing the flow used to match hidden-space shapes and only using a linear mapping for dimension conversion; 4) Decoding strategy comparison: respectively adopting Top 1, Top 2, and Top 4 decoding strategies to train DMIR, and reporting their impact on downstream performance and inference speed on OpenAudioBench. The detailed benchmark sub-categories are provided in Table 3.

4.3 Main Results

Reconstruction and Interpretability As shown in Table 2, the default setting achieves 99.9% Token ACC and 99.5% R_{rel}. These results prove the text VAE provides meaningful latent targets. Appendix I provides a decoded reasoning sample.

Downstream Policy Evaluation As shown in the main results of Table 3 and Figure 2, DeIRD achieves the best performance across all three backbone architectures and the lowest latency among all "think and speak decoupled" models. We observe four findings. **Finding 1:** DeIRD enhances instruction-following capabilities, mitigating context instruction forgetting in traditional explicit thinking: VoiceBench-IFEv.: Step 57.4 (+11.9). **Finding 2:** The magnitude of DeIRD's improvement correlates with the backbone's original capability, yielding more substantial gains for weaker models: OpenAudioBench-Avg: Step 62.6 (+7.5), whereas Kimi 70.8 (+0.6). To assess the stability of this modest gain, we train Kimi-Audio three times; the mean improvement and standard deviation are reported in Appendix E. **Finding 3:** Compared to all explicit text-based "think and speak decoupled" reasoning baselines (DIFFA2 CoT (Free), DIFFA2 CoT, and Explicit CoT), DeIRD alleviates the latency bottleneck of traditional paradigms, achieving TTFA: Step 1.50s (-17.10s). **Finding 4:** DIFFA2 CoT (Free) inserts long CoT without training. We attribute its poor performance to distracting the unadapted model from the user's question. The low TTFA of Prompt (Free) under the current paradigm stems from speaking its reasoning directly, leading to degraded instruction following; details are in Appendix H.

4.4 Ablation and Analysis

Table 1 reveals four findings. **Finding 1: DeIRD's performance gains stem from meaningful latents rather than response postponement.** Inserting random continuous noise yields less downstream improvement than DeIRD and damages style control: Kimi-VStyle-Avg: 1.63 (-1.43), while introducing no distinct latency: Step-TTFA: 0.37s (+0.00s). This proves DeIRD's effectiveness relies on meaningful reasoning latents rather than sequence padding to delay responses, with overall latency primarily controlled by the DMIR generation process.

Finding 2: Reasoning capabilities are robust to parameter changes, but inference latency scales with generated latent count. Downstream reasoning metrics are relatively robust to changes in latent dimension and sequence length: Step-GSM8K-Pass@1: 57.5% (-2.9%) for $d = 768$. However, reasoning latency changes drastically with latent generation count m. Shorter sequences reduce latency, with Step-TTFA at 0.80s (-0.70s) for $m = 16$, whereas longer ones increase it to 2.98s (+1.48s) for $m = 64$. Consistent with this finding, latency analysis shows that DMIR's m sequential mask-predict rounds dominate DeIRD's TTFA, taking 1.39s at $m = 32$ (Appendix H).

Finding 3: The LFA module aligns spatial distributions to protect style control with virtually zero latency overhead. Without the flow used to match hidden-space shapes, directly injecting latents degrades the backbone's paralinguistic abilities: Kimi-VStyle-Avg: 1.54 (-1.64). LFA inclusion aligns spatial shapes to prevent this damage, introducing negligible inference delay.

Finding 4: Larger Top-k decoding strategies accelerate DeIRD but compromise output speech style control. Compared to our default strict *Top 1* mask-predict strategy, employing relaxed decoding (*Top 2* or *Top 4*) accelerates DMIR generation and reduces TTFA: Step-Top4-TTFA: 0.47s (-1.03s). However, this accelerated generation introduces noise into thinking latents, consequentially harming generated speech style control: Step-Top4-VStyle-Avg: 2.88 (-0.27).

Source of Implicit Reasoning Gains. Ablations on Step-Audio-2-mini support instance-specific latent contributions beyond noise or learnable prefixes (Appendix D).

5 Conclusion

We presented DeIRD, a decoupled implicit reasoning framework for spoken dialogue models designed to overcome latency and high training costs of explicit CoT paradigms. DeIRD utilizes a masked prediction model to predict fixed-length continuous latents carrying implicit reasoning, guiding the subsequent response. Experiments demonstrate DeIRD reduces TTFA while improving dialogue capabilities, paralinguistic style control, reasoning diversity, and reasoning depth.

AI USE STATEMENT

In this work, we used generative AI tools for dataset construction and manuscript refinement. Specifically, we utilized Gemini 3.1 Pro to generate explicit CoT sequences, Qwen2.5-Omni for data difficulty calibration, and state-of-the-art TTS models (CosyVoice2, CosyVoice3, and IndexTTS 2.5) to synthesize the audio samples for the LatentReason-40k dataset. Additionally, we used Large Language Models to assist with English phrasing, grammar correction, and LaTeX table formatting of this manuscript. We have not used generative AI tools for generating the core algorithmic methodology, fabricating experimental results, or formulating primary scientific claims, and the rest of the required disclosure tasks are not applicable to this work.

We have carefully reviewed all AI-assisted work. We manually inspected a representative subset of the LLM-generated CoT sequences to verify their logical correctness and format compliance. The AI-synthesized audio was strictly filtered using an ASR system with a 5% Word Error Rate (WER) threshold to guarantee acoustic and semantic fidelity. Furthermore, all AI-assisted text refinements and formatting adjustments were comprehensively reviewed by the authors to ensure scientific accuracy and preserve the original intent. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the rigorous reproducibility of our work, we provide comprehensive details covering the datasets, methodology, and training configurations. The complete construction pipeline, filtering criteria, and distribution statistics for the LatentReason-40k dataset are meticulously documented in Appendix B. We will publicly release the complete LatentReason-40k training dataset, including the generated CoT annotations, subject to applicable privacy and redistribution requirements. The exact execution logic of the DMIR inference process is formalized in Algorithm 1, and the specific latent injection mechanisms for different backbone architectures are explained in Appendix A. Furthermore, the hyperparameters for the three-stage training pipeline are provided in Appendix G.

REFERENCES

Yiming Chen, Xianghu racing Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking LLM-based voice assistants. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.17196.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. Shanks: Simultaneous hearing and thinking for spoken language models. *arXiv preprint arXiv:2510.06917*, 2025a. URL https://arxiv.org/abs/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. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models. *arXiv preprint arXiv:2507.15375*, 2025b. URL https://arxiv.org/abs/2507.15375.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Luigi Daniele and Suphavadeeprasit. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training. *arXiv preprint arXiv:(coming soon)*, 2023. URL https://huggingface.co/datasets/LDJnr/Capybara.

Hanwen Du, Yuxin Dong, and Xia Ning. Latent thinking optimization: Your latent reasoning language model secretly encodes reward signals in its latent thoughts, 2026a. URL https://arxiv.org/abs/2509.26314.

Xuan Du, Qiangyu Yan, Wenshuo Li, Borui Jiang, Changming Xiao, Han Shu, and Xinghao Chen. Thinking-while-speaking: A controlled, interleaved reasoning method for real-time speech generation, 2026b. URL https://arxiv.org/abs/2605.20946.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024. URL https://arxiv.org/abs/2411.13577.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition, 2020. URL https://arxiv.org/abs/1910.11473.

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, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, pp. 22199–22213, 2022. URL https://arxiv.org/abs/2205.11916.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations, 2017. URL https://arxiv.org/abs/1704.04683.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Jijie Li, Li Du, Hanyu Zhao, Bo wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. Infinity instruct: Scaling instruction selection and synthesis to enhance language models, 2025a. URL https://arxiv.org/abs/2506.11116.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. Baichuan-audio: A unified framework for end-to-end speech interaction, 2025b. URL https://arxiv.org/abs/2502.17239.

Yangzhuo Li, Shengpeng Ji, Yifu Chen, Tianle Liang, Haoyu Yang, Junboli, Jun Fang, Lin Li, and Qingyang Hong. Dual-reasoner: Bridging interleaved atomicity and streaming latency via thinking-while-talking. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4081–4105, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.199. URL https://aclanthology.org/2026.findings-acl.199/.

Fanfan Liu and Haibo Qiu. Context cascade compression: Exploring the upper limits of text compression, 2025. URL https://arxiv.org/abs/2511.15244.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model, 2025. URL https://arxiv.org/abs/2505.03739.

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

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.

Yan Rong, Chenxing Li, Dong Yu, and Li Liu. Audiogenie-reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning, 2025. URL https://arxiv.org/abs/2509.16971.

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, Chun Chen, Heyu Chen, Jiawei Li, Lei Li, Menghang Zhu, Peidian Li, Qiying Wang, Sirui Deng, Weimin Xiong, Wenshan Huang, Wenyu Yang, Yilin Jiang, Yixin Yang, Yuanyuan Tian, Yue Ma, Yue Yu, Zihan Zhang, Zihao Yue, Bangjun Xiao, Bingquan Xia, Bofei Gao, Bowen Ye, Can Cai, Chang Liu, Chenhong He, Chunan Li, Dawei Zhu, Duo Zhang, Fengyuan Shi, Guoan Wang, Hailin Zhang, Hanglong Lv, Hanyu Li, Hao Tian, Heng Qu, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianguang Zuo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Linghao Zhang, Meng Chen, Nuo Chen, Peng Zhang, Qianli Chen, Qiantong Wang, Rang Li, Shaohui Liu, Shengfan Wang, Shicheng Li, Shihua Yu, Shijie Cao, Shimao Chen, Shuhao Gu, Weikun Wang, Wenhan Ma, Xiangwei Deng, Xing Yong, Xing Zhang, Xu Wang, Yifan Song, Yihao Zhao, Yingbo Zhao, Yizhao Gao, Yu Cheng, Yu Tu, Yudong Wang, Zhaojun Huang, Zhengju Tang, Zhenru Lin, Zhichao Song, Zhipeng Xu, Zhixian Zheng, and Zihan Jiang. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025. URL https://arxiv.org/abs/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. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025. URL https://arxiv.org/abs/2511.15848.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset, 2024. URL https://arxiv.org/abs/2402.10176.

Guancheng Tu, Xiangjun Fu, Suhao Yu, Yao Tang, Haoqiang Kang, Lianhui Qin, Yizhe Zhang, and Jiatao Gu. Latent reasoning with normalizing flows, 2026. URL https://arxiv.org/abs/2606.06447.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Xia Fei, Chi Ed, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pp. 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

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, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu racing, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025. URL https://arxiv.org/abs/2507.16632.

Donghang Wu, Haoyang Zhang, Jun Chen, Xiangyu, Zhang, Hexin Liu, Eng Siong Chng, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. Mind-paced speaking: A dual-brain approach to real-time reasoning in spoken language models, 2026a. URL https://arxiv.org/abs/2510.09592.

Donghang Wu, Tianyu Zhang, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, and Yoshua Bengio. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning, 2026b. URL https://arxiv.org/abs/2603.17837.

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

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. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, Yifan Peng, Mingquan Lin, and Zongyuan Ge. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding, 2026. URL https://arxiv.org/abs/2603.13366.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, Shimin Li, Jun Song, Xipeng Qiu, and Bo Zheng. Vstyle: A benchmark for voice style adaptation with spoken instructions, 2025. URL https://arxiv.org/abs/2509.09716.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, Yong Qin, and Xuelong Li. Diffa: Large language diffusion models can listen and understand, 2025. URL https://arxiv.org/abs/2507.18452.

Jiaming Zhou, Xuxin Cheng, Shiwan Zhao, Yuhang Jia, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. Diffa-2: A practical diffusion large language model for general audio understanding, 2026. URL https://arxiv.org/abs/2601.23161.

A Latent Injection Mechanism Across Backbone Models

This section details the native generation paradigms of the three backbone models (Step-Audio-2-mini, VITA-Audio, and Kimi-Audio) and exactly how the continuous implicit reasoning latents are inserted into their input contexts:

Step-Audio-2-mini:

- *Native Paradigm*: Single-stream interleaved generation, natively outputting an interleaved sequence of 1 text token followed by up to 4 audio codes.
- *Injection Location*: The mapped latents replace the embeddings of K <dream> placeholder tokens, which are bounded by <|THINK_START|> and <|THINK_END|><tts_start> during the initial forward pass.

VITA-Audio-Plus-Vanilla:

- *Native Paradigm*: Single-stream interleaved generation that outputs a mixed text and audio token stream.
- *Injection Location*: The mapped latents replace K <|fim_pad|> placeholder tokens situated between <think>\n and \n</think>\n in the assistant prefix.

Kimi-Audio-7B-Instruct:

- *Native Paradigm*: Dual-lane parallel generation, which utilizes two separate output heads and a fixed audio delay to generate text and speech streams with coupled states.
- *Injection Location*: The mapped latents replace the complete embeddings of K reserved blank time steps. These placeholders are located between <|reserved_token_0|> and <|reserved_token_1|> in the text lane, and the replacement occurs immediately after the embeddings of the text and audio lanes are merged.

B Details of LatentReason-40k Dataset

The LatentReason-40k dataset comprises a total of 46,809 training instances, amounting to 685.02 hours of high-fidelity audio data. A word cloud of the CoT texts is shown in Figure 4. To ensure comprehensive capability acquisition for the spoken dialogue backbone, the dataset is meticulously divided into two distinct subsets: the Semantic Content subset and the Paralinguistic (Acoustic) subset. Table 4 and Figure 3a Left delineates the detailed distribution of samples and audio durations across all fine-grained task categories.

Semantic Content Subset. The Semantic Content subset (comprising 37,427 instances and 634.30 hours) is curated from 14 diverse open-source corpora (Köpf et al., 2023; Lai et al., 2017; Li et al., 2025a; Rajani et al., 2023; Toshniwal et al., 2024; Yu et al., 2023; LI et al., 2024; Welbl et al., 2017; Jin et al., 2019; Clark et al., 2018; Khot et al., 2020; Daniele & Suphavadeeprasit, 2023). As shown in Figure 3a Right, it encompasses 8 fine-grained task categories: Math, Science QA, Instruction, Logic, Code, Creative, Ultra, and Dialogue.

Paralinguistic Subset. As shown in Figure 3b Left, the Paralinguistic subset (comprising 9,382 instances and 50.73 hours) is specifically designed to endow the model with robust acoustic perception and control capabilities. It encompasses 3 distinct task categories: explicit understanding, explicit generation, and implicit reasoning, comprehensively covering critical acoustic attributes such as emotion, gender, pitch, volume, age, accent, speaking rate, and language.

To bridge the modality gap and transform these heterogeneous text-based semantic queries into a high-quality, speech-centric training corpus, we engineered a rigorous four-stage data processing pipeline. The concrete implementation of each stage is detailed below:

Stage 1: Rule-Based Formatting and Length Filtration. In the initial phase, we deploy heuristic rules and regular expressions to systematically eliminate samples containing structures inherently unsuitable for auditory generation. Specifically, textual instances heavily reliant on visual layouts—such as complex Markdown tables, HTML tags, URLs, JSON formats, or dense programming syntax—are aggressively discarded. Additionally, to emulate the natural pacing of human-machine voice interactions and prevent the synthesis of excessively long audio clips, we impose strict length (0 to 250 words) bounds on the prompts and filter out inputs with an anomalous proportion of non-ASCII characters to maintain linguistic consistency.

Table 3: **Downstream Policy Evaluation and Ablation Studies.** We compare the DeIRD framework against multiple baselines and validate its design choices across four ablation settings. Different background colors represent different spoken dialogue backbone models (blue for Step-Audio-2-mini, orange for VITA-Audio, and green for Kimi-Audio). Within the same experimental setting and model in the Main Results, the best results are **bolded**, and the second best are underlined.

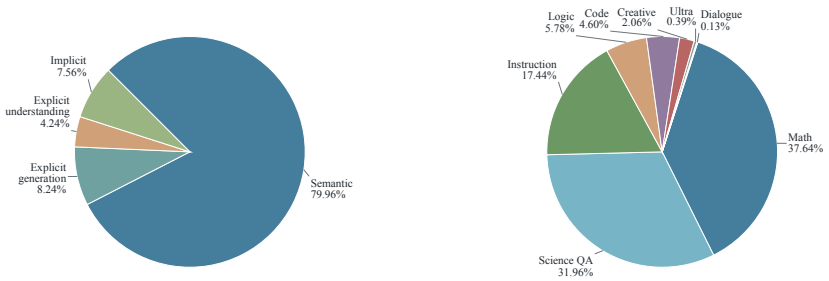
Method	VoiceBench					VStyle					GSM8K				OpenAudioBench					Latency
	Alpaca	OBQA	BBH	IFEv.	Avg	Acous.	Instr.	Role	Emp.	Avg	P@1(%)	P@2(%)	P@4(%)	P@8(%)	Llama	Reas.	Triv.	Web	Avg	TTFA
Main Results: Downstream Policy Evaluation																				
Step-Audio-2-mini	**3.49**	70.1	54.9	42.2	58.7	**3.03**	**3.15**	2.01	4.13	3.08	26.5	44.1	57.6	71.3	66.0	3.03	46.2	51.5	55.1	**0.37**
+ DIFFA2 CoT (Free)	3.05	60.9	55.7	26.0	40.0	1.12	1.19	1.02	1.14	1.12	13.2	19.4	39.1	54.6	52.7	1.32	35.4	31.1	34.4	14.2
+ DIFFA2 CoT	3.03	77.8	56.1	54.4	62.4	2.18	2.11	2.14	3.32	2.44	30.0	52.8	58.4	70.2	72.4	2.90	53.8	58.4	58.2	14.2
+ Prompt (Free)	2.88	76.5	57.3	35.7	60.3	3.07	2.96	2.20	4.08	3.08	27.0	40.9	54.9	68.7	73.8	3.28	53.4	57.2	61.5	0.42
+ Explicit CoT	3.09	76.9	58.1	45.5	62.5	2.44	2.11	2.18	3.43	2.54	32.1	54.3	59.2	78.4	66.9	3.11	46.8	52.5	55.7	18.6
+ DeIRD (Ours)	3.19	**79.6**	**59.2**	**57.4**	**64.9**	2.90	3.07	**2.29**	**4.33**	**3.15**	**60.4**	**71.5**	**80.9**	**88.1**	**74.0**	**3.34**	**54.6**	**58.9**	**62.6**	1.50
VITA-Audio	3.24	64.4	52.5	27.7	50.8	2.22	1.84	1.99	4.01	2.52	64.1	76.1	82.4	88.9	64.5	2.94	38.8	47.0	52.4	**0.11**
+ DIFFA2 CoT (Free)	2.65	44.3	47.2	22.8	44.7	1.03	1.14	1.70	2.98	1.71	20.1	43.8	56.2	60.8	49.0	1.13	37.2	31.1	32.8	12.8
+ DIFFA2 CoT	3.18	64.2	37.7	34.5	53.0	2.31	2.01	2.38	4.29	2.75	50.0	61.5	71.2	78.9	68.0	2.29	40.8	46.8	49.9	12.8
+ Prompt (Free)	3.58	71.2	58.5	30.4	58.9	2.45	2.01	1.91	3.94	2.58	62.8	75.4	82.4	88.8	75.1	2.85	**45.7**	52.5	56.1	0.13
+ Explicit CoT	3.15	**74.3**	**59.7**	39.6	58.7	2.38	2.35	**2.48**	4.02	2.81	59.6	70.3	79.2	85.4	71.1	3.00	43.9	52.0	55.9	9.13
+ DeIRD (Ours)	**3.67**	73.4	53.7	**56.0**	**60.7**	**2.55**	**2.36**	2.41	**4.32**	**2.91**	**65.6**	**77.3**	**86.7**	**91.0**	**75.4**	**3.23**	42.1	**52.9**	**57.6**	1.57
Kimi-Audio-7B-Instr.	**4.09**	83.5	68.4	61.5	72.8	**2.94**	2.69	2.51	4.08	3.06	80.0	88.1	93.6	96.1	81.7	3.36	61.4	69.6	70.2	**0.22**
+ DIFFA2 CoT (Free)	3.35	64.2	43.2	37.8	50.8	1.64	2.06	1.90	3.17	2.19	59.1	72.9	81.2	84.0	65.7	2.80	53.2	49.8	54.9	16.3
+ DIFFA2 CoT	3.64	80.7	65.6	63.8	70.8	2.83	2.01	2.46	4.18	2.87	76.3	84.9	90.1	94.7	77.3	2.87	60.4	63.1	65.6	16.3
+ Prompt (Free)	3.85	83.1	**68.4**	55.2	71.9	2.51	2.32	2.41	4.27	2.88	78.1	88.0	**94.0**	96.1	81.9	3.13	63.0	**70.0**	69.5	0.26
+ Explicit CoT	3.86	**84.6**	63.5	64.1	73.0	2.73	**3.07**	2.69	4.06	3.14	78.2	87.4	92.8	95.9	77.8	3.06	**63.1**	68.0	67.6	12.23
+ DeIRD (Ours)	3.85	83.5	**68.4**	**70.0**	**74.4**	**2.94**	2.50	**2.85**	**4.42**	**3.18**	**80.1**	**88.1**	92.3	**97.2**	**82.9**	**3.42**	**63.5**	68.9	**70.8**	1.67
Ablation 1: Semantic Validity of Latent Variables																				
Step-Audio-2-mini	3.49	70.1	54.9	42.2	58.7	3.03	3.15	2.01	4.13	3.08	26.5	44.1	57.6	71.3	66.0	3.03	46.2	51.5	55.1	0.37
+ Noise Latent	3.24	73.4	56.1	44.7	60.7	2.75	2.38	2.38	3.83	2.84	55.3	69.0	74.6	76.6	69.6	2.94	46.6	54.5	57.7	0.37
VITA-Audio	3.24	64.4	52.5	27.7	50.8	2.22	1.84	1.99	4.01	2.52	64.1	76.1	82.4	88.9	64.5	2.94	38.8	47.0	52.4	0.11
+ Noise Latent	3.63	73.4	50.4	55.6	59.6	2.30	2.30	2.37	4.16	2.78	59.8	72.1	81.1	88.1	76.2	2.87	43.9	52.0	56.3	0.12
Kimi-Audio-7B-Instr.	4.09	83.5	68.4	61.5	72.8	2.94	2.69	2.51	4.08	3.06	80.0	88.1	93.6	96.1	81.7	3.36	61.4	69.6	70.2	0.22
+ Noise Latent	3.76	83.3	58.9	68.0	71.7	2.12	1.75	1.37	1.29	1.63	79.8	88.3	92.5	94.8	82.1	2.95	65.0	67.2	68.0	0.22
Ablation 2: Optimal Parameter Exploration																				
+ DeIRD (Ours)	3.19	79.6	59.2	57.4	64.9	2.90	3.07	2.29	4.33	3.15	60.4	71.5	80.9	88.1	74.0	3.34	54.6	58.9	62.6	1.50
+ DeIRD (2048, 32)	3.19	79.6	59.2	57.4	64.9	2.90	3.07	2.29	4.33	3.15	60.4	71.5	80.9	88.1	74.0	3.34	54.6	58.9	62.6	1.50
+ DeIRD (2048, 64)	2.97	77.6	59.4	55.8	63.8	2.61	2.61	2.33	3.86	2.85	59.0	70.9	80.2	87.3	73.5	3.31	55.0	57.4	62.4	2.98
+ DeIRD (2048, 16)	2.96	74.5	58.4	58.0	63.3	2.56	2.67	2.35	4.03	2.90	58.2	71.6	80.8	88.1	73.7	3.24	55.8	58.7	62.5	0.80
+ DeIRD (3584, 32)	3.03	76.5	58.1	58.1	64.1	2.64	2.49	2.08	3.71	2.73	57.8	70.4	80.5	87.0	73.3	3.25	54.5	56.8	61.6	1.51
+ DeIRD (768, 32)	3.00	76.5	59.6	54.2	63.7	2.67	2.52	2.30	3.99	2.87	57.5	70.6	79.3	86.1	72.8	3.36	53.6	57.7	62.5	1.50
+ DeIRD (Ours)	3.67	73.4	53.7	56.0	60.7	2.55	2.36	2.41	4.32	2.91	65.6	77.3	86.7	91.0	75.4	3.23	42.1	52.9	57.6	1.57
+ DeIRD (2048, 32)	3.67	73.4	53.7	56.0	60.7	2.55	2.36	2.41	4.32	2.91	65.6	77.3	86.7	91.0	75.4	3.23	42.1	52.9	57.6	1.57
+ DeIRD (2048, 64)	3.48	74.1	49.7	55.8	59.9	2.61	2.31	2.33	4.03	2.82	58.7	77.1	81.6	88.0	74.0	2.75	42.6	51.6	55.1	3.03
+ DeIRD (2048, 16)	3.43	75.2	49.7	55.5	59.7	2.42	2.33	2.44	4.06	2.81	60.6	72.6	82.1	87.9	74.8	2.88	42.7	51.7	55.4	0.86
+ DeIRD (3584, 32)	3.53	69.5	49.8	56.5	59.3	2.31	2.44	2.45	4.16	2.84	58.8	71.0	81.9	88.4	75.1	2.91	42.5	52.2	55.4	1.57
+ DeIRD (768, 32)	3.59	72.5	48.4	56.0	60.1	2.33	2.05	2.30	3.96	2.66	59.5	72.9	81.6	88.0	75.3	2.88	43.7	52.1	56.2	1.56
+ DeIRD (Ours)	3.85	83.5	68.4	70.0	74.4	2.94	2.50	2.85	4.42	3.18	80.1	88.1	92.3	97.2	82.9	3.42	63.5	68.9	70.8	1.67
+ DeIRD (2048, 32)	3.85	83.5	68.4	70.0	74.4	2.94	2.50	2.85	4.42	3.18	80.1	88.1	92.3	97.2	82.9	3.42	63.5	68.9	70.8	1.67
+ DeIRD (2048, 64)	3.73	83.5	62.6	66.6	71.5	2.10	1.78	1.58	1.40	1.72	79.2	87.1	92.0	95.3	81.2	2.84	64.2	66.7	67.2	3.16
+ DeIRD (2048, 16)	3.75	83.1	67.7	69.0	72.4	2.85	2.58	2.80	4.31	3.14	81.4	88.9	93.2	95.4	82.1	3.04	64.4	67.0	69.5	0.96
+ DeIRD (3584, 32)	3.93	82.6	65.6	67.8	74.2	2.87	2.59	2.92	4.48	3.22	80.1	87.4	92.3	95.1	80.8	3.15	62.3	67.6	69.4	1.67
+ DeIRD (768, 32)	3.69	82.9	60.2	67.6	71.7	2.88	2.80	2.95	4.54	3.29	80.5	87.5	92.0	95.5	81.6	3.05	64.3	67.1	68.6	1.66
Ablation 3: Necessity of the LFA Module																				
+ DeIRD (Ours)	3.19	79.6	59.2	57.4	64.9	2.90	3.07	2.29	4.33	3.15	60.4	71.5	80.9	88.1	74.0	3.34	54.6	58.9	62.6	1.50
- LFA Alignment	3.01	76.5	59.8	56.6	63.8	2.74	2.43	2.25	3.80	2.81	59.5	71.0	80.2	86.8	70.4	3.28	51.7	55.3	60.3	1.50
+ DeIRD (Ours)	3.67	73.4	53.7	56.0	60.7	2.55	2.36	2.41	4.32	2.91	65.6	77.3	86.7	91.0	75.4	3.23	42.1	52.9	57.6	1.57
- LFA Alignment	3.55	76.0	48.8	53.5	60.0	2.29	2.38	2.42	4.27	2.84	60.5	73.1	82.3	88.6	74.2	2.93	42.9	52.9	56.6	1.55
+ DeIRD (Ours)	3.85	83.5	68.4	70.0	74.4	2.94	2.50	2.85	4.42	3.18	80.1	88.1	92.3	97.2	82.9	3.42	63.5	68.9	70.8	1.67
- LFA Alignment	3.62	85.3	64.8	66.8	72.3	2.02	1.73	1.36	1.06	1.54	79.2	87.0	91.4	95.2	81.4	2.85	64.0	65.4	67.2	1.66
Ablation 4: Decoding Strategy Comparison																				
Top 1 Decoding (Ours)	3.19	79.6	59.2	57.4	64.9	2.90	3.07	2.29	4.33	3.15	60.4	71.5	80.9	88.1	74.0	3.34	54.6	58.9	62.6	1.50
Top 2 Decoding	2.96	76.9	61.5	57.0	64.3	2.58	2.61	2.26	3.92	2.84	56.6	70.2	80.1	87.0	74.1	3.29	54.9	57.3	62.4	0.81
Top 4 Decoding	2.98	76.9	58.3	56.5	64.5	2.48	2.59	2.25	4.20	2.88	59.5	70.4	80.6	88.3	73.8	3.30	54.8	56.8	62.3	0.47
Top 1 Decoding (Ours)	3.67	73.4	53.7	56.0	60.7	2.55	2.36	2.41	4.32	2.91	65.6	77.3	86.7	91.0	75.4	3.23	42.1	52.9	57.6	1.57
Top 2 Decoding	3.51	75.8	49.4	54.0	59.8	2.34	2.32	2.36	4.10	2.78	59.9	72.4	81.3	87.4	75.0	2.94	42.2	52.5	56.6	0.87
Top 4 Decoding	3.55	75.6	49.6	57.4	60.4	2.42	2.31	2.43	4.14	2.83	59.3	72.6	82.2	88.1	74.5	2.78	42.8	52.4	56.1	0.52
Top 1 Decoding (Ours)	3.85	83.5	68.4	70.0	74.4	2.94	2.50	2.85	4.42	3.18	80.1	88.1	92.3	97.2	82.9	3.42	63.5	68.9	70.8	1.67
Top 2 Decoding	3.73	81.1	63.0	68.1	72.6	2.04	1.71	1.56	1.14	1.61	80.5	89.2	92.1	95.6	82.2	2.93	63.6	67.5	68.0	0.97
Top 4 Decoding	3.54	84.2	62.7	69.2	72.4	2.11	1.74	1.57	1.03	1.61	79.9	87.3	92.1	95.0	82.9	3.02	64.4	66.3	63.7	0.62

Stage 2: LLM-Assisted Domain Classification and Modality Adaptation. Subsequent to the rule-based cleansing, Large Language Models (LLMs) are utilized to categorize the surviving semantic prompts into orthogonal domains. Specifically, we employ zero-shot prompting to instruct the LLM to analyze the core semantic intent and task requirements of each query, assigning it to the most appropriate category. This categorical metadata is crucial for guiding subsequent data sampling to maintain domain diversity and a balanced distribution across the final training corpus. Concurrently, a critical filtering criterion enforced during this LLM evaluation phase is "audio-only solvability." Problems that intrinsically depend on visual cues or spatial layout comprehension (e.g., complex geometric figures, nested fractional equations, or spatial grids) are aggressively pruned. We strictly retain only those queries that can be fully comprehended and accurately resolved through pure auditory reception.

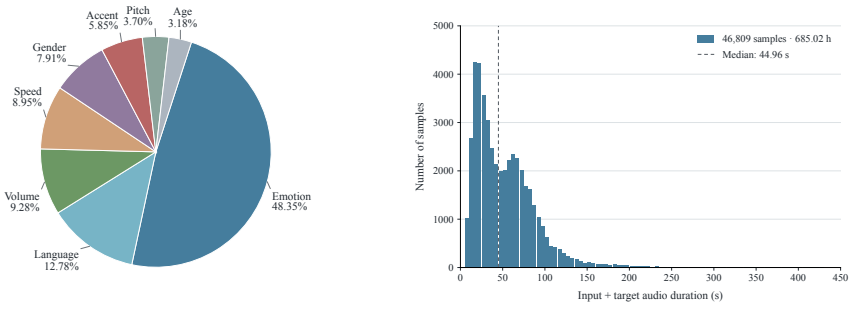
Stage 3: Difficulty Calibration via Qwen2.5-Omni. To ensure the dataset appropriately aligns with the learning capacity of a 7B-parameter spoken dialogue backbone, we introduce a difficulty-

Table 4: Distribution and audio duration of the LatentReason-40k training dataset.

Category	Source Datasets	Samples	Duration (h)
Semantic Content Subset			
Math	Infinity, MetaMathQA, OpenMathInstruct	14,088	291.82
Science QA	ARC, PubMedQA, QASC, SciQ	11,961	163.01
Instruction	No Robots, WizardLM, IF-RL	6,529	63.60
Logic	SlimOrca	2,162	39.06
Code	Infinity, OpenAssistant	1,722	62.17
Creative	Airoboros, Infinity	770	11.14
Ultra	Capybara, Chatbot Arena, Pure-Dove, UltraChat	146	2.95
Dialogue	OASST1	49	0.54
Paralinguistic Subset			
Explicit Generation	–	3,856	23.53
Explicit Understanding	–	1,987	6.10
Implicit	–	3,539	21.09
Total	**–**	**46,809**	**685.02**



(a) **Left:** Overall distribution of LatentReason-40k; **Right:** Task of the semantic subset.



(b) **Left:** Attribute of the paralinguistic subset; **Right:** Distribution of audio duration.

Figure 3: Visualization of the LatentReason-40k training dataset.

aware evaluation mechanism. Employing Qwen2.5-Omni (Xu et al., 2025) as the evaluator, we gauge the empirical complexity of each candidate sample. For each prompt, we generate 8 independent responses using stochastic sampling (`do_sample=True`) at a temperature of 1.0. We calculate the empirical accuracy (p_{corr}) based on these 8 trials and strictly retain only moderate-difficulty samples whose correctness probability falls within the interval of $[0.125, 0.875]$, thereby effectively discarding trivial queries and unsolvable anomalies.

Table 5: Reconstruction performance and query relevance across different representation and decoding combinations.

Representation	Decoding	Text Recon.		Query Relevance
		Token ACC (↑)	WER (↓)	R_{rel} (↑)
AE	Flow Matching	99.5%	0.015	12.5%
VAE	Flow Matching	99.9%	0.003	49.0%
VAE (DeIRD)	**Mask Predict**	**99.9%**	**0.003**	**99.5%**

Table 6: Downstream policy evaluation on OpenAudioBench using the Step Audio 2 mini backbone to compare decoding strategies.

Method	OpenAudioBench				
	Llama	Reas.	Triv.	Web	Avg
Baseline	66.0	3.03	46.2	51.5	55.1
+ VAE and Flow Matching	67.2	2.88	44.7	52.5	55.8
+ DeIRD (VAE and Mask Predict)	**74.0**	**3.34**	**54.6**	**58.9**	**62.6**

Stage 4: Audio Synthesis and Structured CoT Generation. In the final stage, the meticulously curated textual queries are converted into high-fidelity speech utilizing an ensemble of state-of-the-art text-to-speech (TTS) systems, specifically CosyVoice2, CosyVoice3, and IndexTTS 2.5. To maximize acoustic diversity, we extract reference speaker timbres directly from the Seed-TTS-Eval benchmark. Following the synthesis, we implement a rigorous post-generation quality filter: an Automatic Speech Recognition (ASR) model is employed to transcribe the generated audio, and any samples exhibiting a Word Error Rate (WER) exceeding 5% or unnatural acoustic artifacts are aggressively discarded, thereby guaranteeing the acoustic clarity and semantic fidelity of the inputs. Concurrently, to construct the explicit reasoning targets required for text VAE compression, we prompt Gemini 3.1 Pro to generate corresponding Chain-of-Thought (CoT) sequences for all samples. To guarantee optimal fixed-length compression by the VAE module, the maximum CoT length is strictly capped at 1,500 words. Furthermore, all generated CoTs strictly adhere to a unified, three-part structured template: (1) an analysis of the paralinguistic traits of the input audio *(e.g., "the user spoke in an angry tone")*, (2) a rigorous logical deduction addressing the semantic content of the query, and (3) acoustic paralinguistic specifications for the target output *(e.g., "a gentle tone is required for the response")*.

C Exploration of Latent Representations and Decoding Strategies

To encode explicit reasoning sequences into fixed length continuous variables, text compression utilizes either Autoencoder (AE) or Variational Autoencoder (VAE) representations. Since DMIR generates continuous targets, Flow Matching serves as a standard decoding alternative. The main text adopts VAE representation combined with Mask Predict decoding. This section explores three alternative architectural combinations.

C.1 Representation Initialization and Flow Matching Formulation

AE Representation and Initialization. The deterministic AE architecture is initialized directly from the pretrained C3 Text Autoencoder weights. It utilizes a text encoder to process the input sequence and extracts hidden states h_{query} at m predefined learned query positions. A linear projector subsequently maps these hidden states to continuous latent slots $z_{1:m}$. These deterministic latents $z_{1:m}$ are then directly fed into the causal decoder to auto regressively reconstruct the original text.

Algorithm 1 DMIR Inference via VAE and Mask-Predict

Input: *User audio* p, *Mask token* M_{token}, *latent dimension* $d = 2048$
Output: *Implicit latent sequence* $\hat{z}_{1:m} \in \mathbb{R}^{m \times d}$
 # 1. Initialize mask status and empty continuous latents
$M \leftarrow \mathbf{1}^{m \times 1}$
$\hat{z}_{1:m} \leftarrow \mathbf{0}^{m \times d}$
for round $= 1$ **to** m **do**
  # 2. Match hidden-space shapes and dimensions via Latent Flow Adapter
  $h_{\text{vis}} \leftarrow \text{LFA}(\hat{z}_{1:m})$ *# Project to unified hidden state space*
  # 3. Construct current visible slot features z_{vis}
  for $k = 1$ **to** m **do**
    if $M[k] == 1$ **then**
      $z_{\text{vis}}[k] \leftarrow M_{\text{token}}$
    else
      $z_{\text{vis}}[k] \leftarrow h_{\text{vis}}[k]$
    end if
  end for
  # 4. Predict Gaussian parameters for all masked slots in parallel
  $\hat{\mu}_{1:m}, \log \hat{\sigma}_{1:m} \leftarrow g_\theta(p, z_{\text{vis}}, M)$ *# g_θ represents DMIR*
  $c_k \leftarrow -\frac{1}{d} \sum_{j=1}^{d} \log \hat{\sigma}_{k,j}$ ($\forall k$ where $M[k] == 1$)
  # 5. Retain the uncommitted slot with the highest confidence
  $k^* \leftarrow \text{argmax}_{k|M[k]=1}(c_k)$
  $\hat{z}_{k^*} \leftarrow \hat{\mu}_{k^*}$
  $M[k^*] \leftarrow 0$
end for
return $\hat{z}_{1:m}$

Flow Matching Decoding. Flow matching constructs a continuous probability path from noise to data. The forward process is defined as:

$$z_t = (1 - t)z_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad t \in [0, 1] \tag{14}$$

The model predicts the velocity v_{pred} optimized via Mean Squared Error against the target $v_{\text{target}} = \epsilon - z_0$. During inference, the latent is decoded using Euler integration:

$$z = z - (t_{\text{current}} - t_{\text{next}})v_{\text{pred}} \tag{15}$$

C.2 Architectural Combinations

AE Representation with Mask Predict. The Mask Predict mechanism strictly requires variance output to compute confidence scores for iterative slot unmasking. Since the deterministic AE representation lacks variance estimation, this combination provides no confidence metric and is architecturally infeasible.

AE Representation with Flow Matching. As shown in Table 5, AE representation achieves accurate text reconstruction. However, the deterministic continuous latent space lacks smoothness. When Flow Matching predicts velocity trajectories across this non smooth space, it causes significant deviations. Consequently, the decoded latents exhibit poor query relevance.

VAE Representation with Flow Matching. VAE representation combined with Flow Matching yields good reconstruction quality. However, its decoded texts show lower relevance to input queries than those from the Mask Predict approach. Downstream evaluation on Step Audio 2 mini across OpenAudioBench (Table 6) demonstrates that reasoning capabilities under this setting are inferior to the standard DeIRD framework. These results support the effectiveness of combining VAE representations with Mask Predict for downstream reasoning.

D Ablation Studies on the Source of Implicit Reasoning Gains

To demonstrate that DeIRD's performance gains originate from actual implicit reasoning based on semantically aligned latents, we must disentangle these gains from the structural artifacts of delayed generation, extra trainable parameters, or mere backbone adaptation. Therefore, we conduct a comprehensive set of ablation studies to prove that the current performance cannot be trivially obtained through delayed responses (noise) or structurally similar soft-prompting techniques.

Table 7: Downstream policy evaluation on VoiceBench and OpenAudioBench using the Step-Audio-2-mini backbone to isolate the source of implicit reasoning gains. The best results are **bolded**, and the second best are underlined.

Method	VoiceBench					OpenAudioBench				
	Alpaca	OBQA	BBH	IFEv.	Avg	Llama	Reas.	Triv.	Web	Avg
Baseline	**3.49**	70.1	54.9	42.2	58.7	66.0	3.03	46.2	51.5	55.1
DeIRD (Ours)	3.19	**79.6**	**59.2**	**57.4**	**64.9**	74.0	**3.34**	**54.6**	**58.9**	**62.6**
+ Noise Latent	3.24	73.4	56.1	44.7	60.7	69.6	2.94	46.6	54.5	57.7
+ Constant Latent	2.98	72.7	55.7	55.2	60.4	73.2	3.28	53.7	52.3	59.1
+ Shuffled Latent	3.04	76.3	50.1	55.9	60.4	**74.6**	3.23	50.4	54.1	58.8
+ Learnable Prefix	2.97	77.1	55.6	57.1	61.3	71.8	3.30	50.5	55.2	58.9
+ Frozen LFA	2.97	74.7	54.8	57.2	61.0	73.3	3.27	48.6	55.9	58.7
+ Eval w/o Latent	3.06	77.6	56.9	56.2	60.8	72.7	3.15	51.3	51.6	57.8

Experimental Settings. To isolate the precise source of performance improvements, we evaluate six distinct contrastive setups on the Step-Audio-2-mini backbone: 1) **Noise Latent**: replacing DMIR latents with random noise sampled from the expected continuous space; 2) **Constant Latent**: replacing instance-specific latents with a single set of fully learnable, shared latent parameters, eliminating semantic variance across samples; 3) **Shuffled Latent**: deliberately destroying the audio-to-latent correspondence by assigning DMIR latents from mismatched samples during inference; 4) **Learnable Prefix**: bypassing the LFA module entirely and directly inserting trainable embeddings of matching dimensions to emulate standard soft-prompt tuning; 5) **Frozen LFA**: freezing the LFA parameters during downstream fine-tuning while using valid DMIR latents; 6) **Eval w/o Latent**: training the backbone with normal DMIR latents but removing them entirely during inference.

Analysis. The results in Table 7 unequivocally confirm that meaningful semantic alignment is the driver of DeIRD's reasoning improvements. While techniques introducing additional trainable tokens or simply delaying responses (e.g., Learnable Prefix, Constant Latent, Noise Latent) marginally improve baseline performance due to extra model capacity or compute time, they consistently fall short of DeIRD (OpenAudioBench Avg: DeIRD 62.6 vs. Learnable Prefix 58.9). Furthermore, destroying the instance-specific semantic mapping (Shuffled Latent) or removing the latents during inference (Eval w/o Latent) causes capability drops. This proves that DeIRD's success stems from injecting correct and instance-specific reasoning semantics rather than relying on soft-prefix or delayed-generation artifacts.

E Training Stability on Kimi-Audio

To assess the stability of DeIRD's modest gains on Kimi-Audio-7B-Instr., we conduct three training runs using the same training data, hyperparameters, GPU configuration, and evaluation protocol, varying only the training random seed. Table 8 reports individual results and their mean and sample standard deviation.

Across three training runs, DeIRD achieves an OpenAudioBench average of 70.80 ± 0.10, with every run exceeding the baseline of 70.2. VoiceBench and VStyle averages also improve consistently, reaching 74.37 ± 0.15 and 3.18 ± 0.02, compared with baseline scores of 72.8 and 3.06, respectively. These results support consistent gains across the three runs, although improvements are not uniform across individual metrics.

Table 8: **Training stability on Kimi-Audio-7B-Instr.** Results of three DeIRD training runs. The final two rows report the mean and sample standard deviation across the three runs, respectively.

Method	VoiceBench					VStyle					GSM8K				OpenAudioBench				
	Alpaca	OBQA	BBH	IFEv.	Avg	Acous.	Instr.	Role	Emp.	Avg	P@1(%)	P@2(%)	P@4(%)	P@8(%)	Llama	Reas.	Triv.	Web	Avg
Kimi-Audio-7B-Instr.	4.09	83.5	68.4	61.5	72.8	2.94	2.69	2.51	4.08	3.06	80.0	88.1	93.6	96.1	81.7	3.36	61.4	69.6	70.2
DeIRD 1	3.85	83.5	68.4	70.0	74.4	2.94	2.50	2.85	4.42	3.18	80.1	88.1	92.3	97.2	82.9	3.42	63.5	68.9	70.8
DeIRD 2	3.82	83.3	68.2	69.8	74.2	2.92	2.48	2.84	4.40	3.16	80.2	88.0	92.2	97.0	82.7	3.40	63.3	68.8	70.7
DeIRD 3	3.87	83.6	68.5	70.2	74.5	2.95	2.52	2.87	4.44	3.20	80.3	88.3	92.5	97.3	83.0	3.43	63.6	69.1	70.9
Mean	3.85	83.47	68.37	70.00	74.37	2.94	2.50	2.85	4.42	3.18	80.20	88.13	92.33	97.17	82.87	3.42	63.47	68.93	70.80
Std	0.03	0.15	0.15	0.20	0.15	0.02	0.02	0.02	0.02	0.02	0.10	0.15	0.15	0.15	0.15	0.02	0.15	0.15	0.10

F DMIR Inference Pseudo-code

This section provides the pseudo-code for the DeIRD implicit reasoning phase. As detailed in Algorithm 1, it illustrates the exact step-wise execution logic of DMIR using VAE representations and Mask-Predict decoding, clarifying how the model iteratively unmasks latent slots via variance-derived confidence, transforms features to match hidden-space shapes, and outputs the final continuous distributions. Specifically, to clarify the dimensional transformations, first, the posterior distribution of the target provided by the text VAE is modeled as a diagonal Gaussian, where μ and $\log \sigma^2$ have a shape of $m \times d$ (with the canonical latent dimension $d = 2048$), allowing the KL divergence to be computed analytically across independent dimensions. Second, since DMIR's internal hidden dimension is 4096, the LFA module initially projects the inputs into the 4096-dimensional space. Subsequently, the raw 4096-dimensional vectors output by $g_\theta(p, z_{\text{vis}}, M)$ are mapped back to 2048 dimensions via a linear projection, and the confidence calculation is performed within this 2048-dimensional space.

G Training Pipeline and Hyperparameters

Table 11 details the hyperparameters for DeIRD. The training consists of three stages:

Stage 1: C3-VAE. The model is initialized from the deterministic C3 VAE. We train the full C3-VAE on 401,019 CoT texts to compress text up to 2048 tokens into 32 continuous posterior means of dimension 2048. The objective is reconstruction Cross-Entropy (CE) combined with a KL penalty weighted at 1e-4. The model is optimized using AdamW with a peak learning rate of 1e-5, a cosine schedule, and a 0.01 warm-up ratio.

Stage 2: Masked DMIR. This stage utilizes 39,401 spoken instructions and their offline VAE means. The training involves an initial 1-epoch warm-up of the Latent Flow Adapter (LFA) using AdamW (learning rate 1e-4, weight decay 0). Subsequently, the LFA is frozen, and we train the semantic and acoustic connectors, DIFFA-2 LoRA, feedback projector, and Gaussian heads for 3 epochs, while keeping the Whisper encoder and LLaDA base frozen. The objective combines masked Mean Squared Error (MSE) and a 0.1-weighted Gaussian Negative Log-Likelihood (NLL). To prevent numerical degeneracy where the NLL could drive the predicted standard deviation $\hat{\sigma} \to 0$ when residuals approach zero, the network predicts the unconstrained log-scale $s_{k,j}$, and we apply a strict lower bound clamp: $\log \hat{\sigma}_{k,j} = \max(s_{k,j}, -5.0)$. This enforces a minimum variance of e^{-10}, ensuring mathematical stability for the variance-derived confidence calculation. We apply constant group-wise learning rates without warm-up.

Stage 3: S2ST SFT. The offline DMIR predictions are integrated into the Step-Audio-2-Mini, Kimi-Audio-7B-Instr, or VITA-Audio base models. Using 46,809 spoken-dialogue examples, we perform full-model supervised fine-tuning alongside the downstream LFA module. The objective is the native mixed text/audio token CE. Training lasts for 1 epoch using AdamW with a 1e-6 learning rate and 0.1 weight decay. Step-Audio utilizes a constant learning rate schedule with ZeRO-3, whereas Kimi-Audio and VITA-Audio use a 0.08 warm-up ratio with ZeRO-2.

Clarification on Training Cost. We explicitly distinguish between the total framework construction cost and the downstream backbone adaptation cost. For framework construction, C3-VAE was trained on 401,019 samples requiring 294 GPU-hours (32 GPUs, 9.2 hours), and DMIR was trained on 39,401 samples requiring 9.3 GPU-hours (8 GPUs, 1.2 hours). The total upfront framework

Table 9: Audio Duration Statistics ((s)) on OpenAudioBench

Method	Mean(s)	P50(s)	Std (σ)	P90(s)	P99(s)
Prompt (Free)	31.95	27.08	14.85	65.47	95.50
DeIRD ($m = 32$)	10.24	7.86	11.64	16.02	31.88

Table 10: **TTFA Breakdown for DeIRD** on Step-Audio-2-mini.

Pipeline Stage	Latency (s)	Description
1. Audio/Context Encoding	0.0725	Whisper encoder extracting features
2. DMIR Inference ($m = 32$)	1.3896	Confidence-guided iterative Mask-Predict
3. LFA Projection	0.0017	Affine-coupling flow and linear mapping
4. Backbone Prefill	0.0236	LLM context prefilling
5. First Audio Generation	0.0198	Generating the first audio token
Total TTFA (Mean)	**1.5071**	Cumulative Time to First Audio

preparation costs approximately 303.3 GPU-hours. In contrast, the downstream backbone adaptation on 46,809 samples is efficient, consuming only 15.6, 14.6, and 11.7 GPU-hours for Step-Audio, VITA-Audio, and Kimi-Audio, respectively. Thus, while the multi-stage pipeline involves initial overhead, endowing any spoken dialogue model with implicit CoT capabilities strictly requires only 1 extra epoch of backbone adaptation.

H Detailed TTFA Protocol and Latency Breakdown

To ensure rigorous latency evaluation, we detail our measurement protocol and provide a micro-architectural breakdown of DeIRD's inference latency.

Measurement Protocol & Hardware Specifications. All latency evaluations are conducted on a single NVIDIA A100 80GB GPU. We evaluate the complete OpenAudioBench dataset with a batch size of 1, executing across 3 independent runs to ensure statistical stability.

Analysis of the Prompt (Free) Baseline. As observed in Table 3, the zero-shot *Prompt (Free)* baseline exhibits an anomalously low TTFA (e.g., 0.42s on Step-Audio), almost identical to the non-reasoning Base Model (0.37s). This occurs because prompting the model to reason causes the reasoning steps to be generated directly via the speech modality. Consequently, while it achieves low latency, it includes excessive reasoning descriptions within the formal spoken response, resulting in two severe drawbacks: prolonged audio outputs and unstable instruction-following capabilities. First, as shown in Table 9, the speech output becomes excessively long: Prompt (Free) averages 31.95s, which is 21.71s longer than DeIRD. Second, mixing reasoning processes into the final audio disrupts instruction execution. As evidenced by the VoiceBench IFEv. metrics in Table 1, Prompt (Free) causes fluctuations in instruction-following capabilities across different backbones.

Micro-Architectural Latency Breakdown. The TTFA of DeIRD is composed of five sequential computational stages. Table 10 reports the average latency breakdown of the Step-Audio-2-mini backbone equipped with DeIRD ($m = 32$) over the complete OpenAudioBench dataset.

Statistical Dispersion. To comprehensively illustrate the stability of DeIRD's latency, we report the distribution statistics of the TTFA evaluated across 3 independent runs on the complete OpenAudioBench dataset. The latency remains highly stable, dominated by the deterministic $m = 32$ iterations of DMIR:

- **Mean:** 1.50s
- **Median (P50):** 1.48s
- **Standard Deviation** (σ)**:** 0.12s

Table 11: Key hyperparameters of the three-stage training pipeline for DeIRD.

Parameter	Stage 1: C3-VAE	Stage 2: Masked DMIR	Stage 3: S2ST SFT
Initialization	C3 VAE	DIFFA-2	Step / Kimi / VITA
Training examples	401,019	39,401	46,809
Latent shape	32×2048	32×2048	32×2048
Trainable modules	Full C3-VAE	LoRA & LFA	Full speech model & LFA
Precision	BF16 + TF32	BF16 (LFA in FP32)	BF16
Optimizer	AdamW (β=0.9, 0.999)	AdamW (β=0.9, 0.999)	AdamW (β=0.9, 0.95)
Weight decay	0	0	0.1
Learning rate	1×10^{-5}	5×10^{-6}-1×10^{-4}	1×10^{-6}
LR schedule	cosine	constant	constant
Warm-up	0.01 ratio	none	0.08 ratio
GPUs	32	8	8
Batch per GPU	10	7	1
Gradient accumulation	2	1	4 / 1 / 1
Effective batch	640	56	32 / 8 / 8
Gradient clipping	1.0	1.0	1.0
Max Sequence Length	2048	32	4096
Selected checkpoint	epoch 2	epoch 3	epoch 1
Distributed strategy	ZeRO-3	ZeRO-3	ZeRO-3 / ZeRO-2 / ZeRO-2

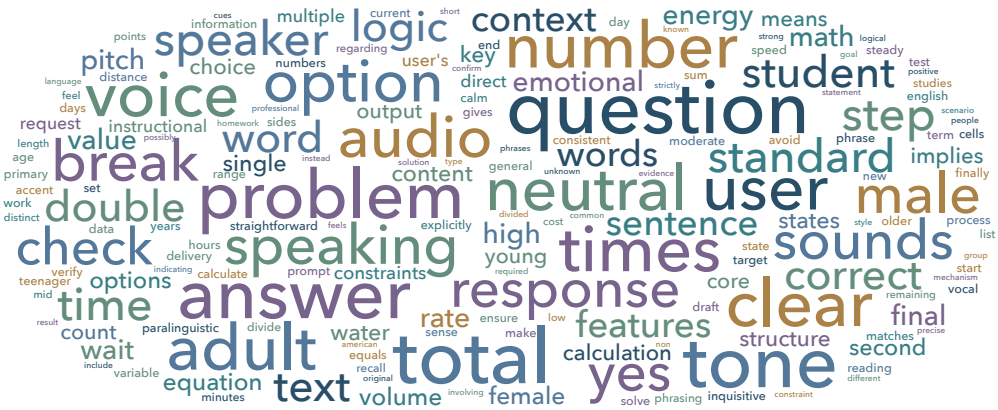


Figure 4: Word cloud of the CoT texts in the LatentReason-40k dataset.

- **90th Percentile (P90):** 1.65s
- **99th Percentile (P99):** 1.78s

I Case Study: Decoded Implicit Latents

Figure 6 presents an end-to-end inference case study of the DeIRD framework. This example demonstrates that the fixed-length implicit continuous latents $\hat{z}_{1:m}$ generated by the DMIR module can be projected back into the discrete text space via the pre-trained text VAE decoder, providing a view into the model's internal thought process.

As shown in the decoded reasoning sequence, the direct text realization of continuous latents naturally exhibits broken grammar, syntactic fragmentation, and occasional token repetition. This is a characteristic phenomenon when mapping high-density continuous embeddings back to discrete tokens. However, despite these grammatical imperfections, the decoded text retains semantic content **relevant** to the user's input. The unified implicit latents $\hat{z}_{1:m}$ successfully capture the complete three-stage reasoning pipeline: recognizing the paralinguistic acoustic cues of the user, performing the core abstract logic deduction necessary to resolve the query, and accurately encapsulating the acoustic delivery planning at the sequence's conclusion.

Prompt 1: Explicit CoT Generation (Understanding & Reasoning)

```
You are generating the INTERNAL chain-of-thought that a
speech-language model would think BEFORE answering a user's spoken
question. You are given the user's spoken question as audio. A
reference transcript of the question is also provided to help you
understand the content accurately. Write the chain-of-thought in
ENGLISH, in first person, as a natural internal monologue (like
"Okay, the user is asking...", "Let me think...", "Wait, let me
double-check..."). Requirements:
1. Begin by briefly noting the paralinguistic cues you perceive
from the AUDIO: the speaker's likely gender, approximate age, and
emotional tone / speaking manner. Keep this to 1-2 sentences and
infer who the user might be and what they want.
2. Then reason through the actual CONTENT of the question step by
step to reach a correct, helpful answer. For math / code / logic,
show the actual working and verify it. This content reasoning is
the main body.
3. Naturally arrive at the answer within the monologue.
Keep the entire monologue concise and self-contained. STRICT
LENGTH LIMIT: do NOT exceed 800 words, and aim for around 400 words.
Output ONLY the monologue text.
Reference transcript of the user's question: """{text}"""
```

Prompt 2: Generation Guidance CoT Generation (Acoustic Delivery)

```
You are the SPEECH model itself, writing a SHORT private note
that plans HOW you will SPEAK your reply out loud. This is the
paralinguistic GENERATION-side thought. It is about VOICE DELIVERY
only (tone, emotion, pacing, energy, pitch, accent, volume) about
the question's meaning or the answer's facts.
You are given: the user's spoken question as AUDIO, your earlier
understanding-side analysis, and the exact reply you are about
to speak. This is an ordinary content question. Plan a natural,
fitting speaking voice for delivering your reply. Do NOT mention
any specific gender, accent, age, or speaker identity
IMPORTANT | instruction following: if the spoken question contains
an EXPLICIT text-level instruction to produce specific words (for
example: "count from one to ten", "read/repeat the following
sentence", "spell this word", "say it three times"), then your note
MUST contain BOTH of these, together: (1) plainly state the strict
instruction-following task | name exactly what must be said | and
that it must be followed precisely with no extra interpretation;
AND (2) the paralinguistic delivery to use. A request such as
"count from one to ten in a child's voice" therefore needs both
the counting task AND the child-like voice. If there is no such
explicit instruction, ignore this paragraph completely and describe
only the delivery. Write in ENGLISH, first person, as one natural
flowing thought. LENGTH: around 40 words, and NEVER exceed 64
words.
Output ONLY the note text | no preamble, no quotes, no labels. Your
earlier understanding-side analysis: {cot}
The reply you are about to speak aloud: {answer-text}
```

Figure 5: Prompts utilized for generating the explicit CoT texts in the LatentReason-40k dataset construction. **Top**: The prompt designed to generate the reasoning trajectory, encompassing paralinguistic analysis and semantic deduction. **Bottom**: The prompt tailored to generate the acoustic delivery planning. During data construction, the outputs from these two distinct generation stages are sequentially concatenated to form the final unified CoT sequence.

[User Spoken Query] (Audio Transcript)

Reactants may also be present in such low concentrations that it is unlikely they will meet and do what? A. explode. B. collide. C. depart. D. melt.

[DeIRD Thinking Phase] Decoded Implicit Reasoning Sequence ($\hat{z}_{1:m}$)

```
Okay hearing user asking chemistry multiple-choice regarding
concentration.  Voice sounds young male neutral, seeking
confirmation tone.  Lower concentration means spatial heavily
dispersed volume.  If sparse fundamentally physical fails happen.
Kinetics mechanism molecules must strike to react.  Explode,
collide, depart, melt.  Mass melting or energetic explode requires
heat transfer, not just spatial.  Core semantic implies meeting
actively is striking.  Scientifically universally particles need
to touch across fluid.  Highly dilute probability drops.  Liquid
gas phase inherently proportional collision frequency.  Fail to
strike implies fail to collide.  Particles like a universally
between passive cause scientifically molecules.  Scientifically that
particles hold carrier force being into electromagnetic substances
waves.  Photons occurring molecules properties implies almost
strongly dissolved less liquid substance water.  Actively absolutely
just strong mechanism implies the mass causing occurring particles
affect extremely mass molecules.  Particles affect usually largely
completely passive, imply inherently intrinsic cause scientifically
molecules compounds.  Inherently force cause into electromagnetic
charge waves.  Particles signals mechanical stems from force
signals dissolved electromagnetic particles, substance water fuels
act.  Mass prevents nit energy forces particles, actively creates
inherently strong mechanism releases an heat compounds.  Particles
sustain happens largely is organic trigger less intrinsic cause
releases passive compounds mechanism.  I will speak clear measured
pace, gentle factual instructional tone.  Consistent calm voice
helpful, no dramatic pitch just steady delivery.
```

[DeIRD Responding Phase] Generated Spoken Response

In chemical kinetics, the fundamental rate-determining step is the physical interaction between molecules, which scientists describe as a collision. When reactants exist at very dilute levels, the probability of them encountering one another becomes astronomically low because the particles are spread thinly throughout the reaction mixture. This specific threshold condition directly controls how quickly the transformation occurs, making the meeting of the reactants the definitive limiting factor rather than an energetic impact or a phase change. The correct selection is option B, collide.

[Reference] Ground Truth Spoken Response

When you consider how chemical reactions actually begin, it all comes down to particles physically bumping into one another. If reactants are highly diluted, those molecules are spread incredibly far apart within the container. Because they are so scattered, the probability of them running into each other drops significantly. Scientists refer to this essential contact as a collision, and without it, no reaction can occur. That makes collide the right choice, corresponding to option B.

Figure 6: An end-to-end inference case study of the DeIRD framework. The fixed-length implicit reasoning latents $\hat{z}_{1:m}$ predicted by DMIR in the continuous space are decoded back into discrete text via the pre-trained text VAE decoder, demonstrating the internal thought process (including paralinguistic analysis, logical deduction, and delivery planning).