DIFFUSPEECH: BIDIRECTIONAL CO-REFINEMENT OF REASONING AND SPEECH VIA UNIFIED SPEECH-TEXT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken language models increasingly reason in text before they speak, but they decode autoregressively, so a reasoning token can never condition on the part of the spoken answer that follows it. We study *Bidirectional Co-Refinement*, in which a text reasoning trace and a spoken reply are denoised jointly by one masked diffusion model: still-masked positions of each modality are predicted from the committed tokens of both, so reasoning can be committed after—and conditioned on—reply speech that follows it. Committed tokens are never rewritten. We build DIFFUSPEECH, a masked-diffusion language model for speech understanding and generation (ASR, TTS, S→T, S→S, T→T) that is, to our knowledge, the first to denoise a reasoning trace jointly with a spoken reply, and THINKINGTALK, a synthetic 26K-sample (319 h) corpus that pairs spoken questions with written reasoning traces and spoken answers. DIFFUSPEECH (8B) reaches 50.6% average S→S accuracy on three spoken-QA benchmarks, above the general-purpose speech LMs we evaluate except Step-Audio 2 mini, and preserves its base LLM's text abilities. Probing the mechanism, we find that most of the benefit of reasoning flows from the trace to the speech. Interventions on the reverse direction—freezing the trace or blocking its attention to speech at inference, matched-training ablations at 1.7B, and error injection—consistently favor joint denoising, with smaller gains concentrated on multi-hop questions (-2.8 points on average and -4.8 on WebQuestions when the trace is frozen before speech). Each probe carries a confound, which we report, so we regard the reverse-direction gain as consistent but not yet conclusive.

1 INTRODUCTION

Chain-of-thought reasoning (Wei et al., 2023; Guo et al., 2025) has reached spoken language models. Autoregressive (AR) speech LMs now place a text reasoning trace between a spoken question and a spoken answer: STITCH (Chiang et al., 2026) alternates unspoken reasoning chunks with spoken chunks, Mini-Omni-Reasoner (Xie et al., 2025) interleaves reasoning and speech tokens, SHANKS (Chiang et al., 2025) reasons while the user is still speaking, and Mind-Paced Speaking (Wu et al., 2025c) lets a reasoning stream pace a separate speaking stream. Intermediate reasoning improves spoken answers, and reasoning and speech can overlap in time at low latency.

These models share one property of causal decoding: tokens are committed in output order, so a reasoning token conditions only on speech that precedes it. In a think-then-speak model, reasoning sees no reply speech at all; interleaved and dual-stream designs let it see earlier speech, but never the part of the answer that comes after it, such as the answer entity the reasoning is working toward. Masked diffusion language models (Sahoo et al., 2024; Nie et al., 2025) remove this coupling: they commit positions in any order, and every prediction conditions on everything committed so far. When a reasoning trace and a spoken reply are denoised *jointly*, confident reply-speech tokens can be committed early and enter the context of reasoning positions that precede them, while committed reasoning conditions the remaining speech. We call this regime **Bidirectional Co-Refinement** (Figure 1). The refinement acts on the predictive distributions of still-masked positions; committed tokens are never rewritten (Section 3.4).

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.

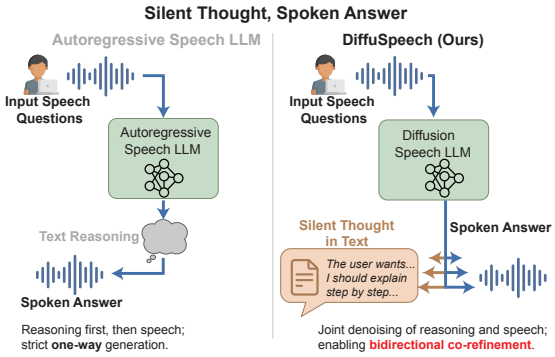

Figure 1: **Commitment order. Left:** an AR model commits its reasoning before any reply speech exists; interleaved AR designs alternate chunks, but reasoning still sees only speech that precedes it. **Right:** DIFFUSPEECH denoises reasoning and reply jointly: still-masked positions of either modality are predicted from all committed tokens of both, so reasoning can be committed after, and conditioned on, reply speech that follows it. Committed tokens are never rewritten.

We ask two questions: can a masked diffusion LM cover the full speech-text stack, and does the reverse, speech-to-reasoning channel contribute? To answer them we build DIFFUSPEECH, a speech-text diffusion LM initialized from LLaDA-8B (Nie et al., 2025), which represents speech as discrete HuBERT units (Hsu et al., 2021) and requires, beyond vocabulary extension, modality-specific masking schedules, selective-masking conditional training, and a delimiter protocol. We also construct THINKINGTALK, a synthetic corpus in which each sample pairs a spoken question, a written reasoning trace, and a spoken answer.

DIFFUSPEECH is competitive with general-purpose AR speech LMs of similar size—50.6% average S→S accuracy, second to Step-Audio 2 mini (52.2%) in our harness—and preserves LLaDA's text abilities. Probing the mechanism (Section 5), we find that the forward channel carries most of the benefit: blocking the reply from attending to the trace erases the gain from reasoning. For the reverse channel, freezing the trace before speech (-2.8) or blocking its attention to speech (-2.7) at 8B, training and decoding the 1.7B model natively in a think-first or attention-restricted regime (26.8 and 25.5 vs. 29.1), and blocking the channel during error recovery (41% vs. 64% recovery) all favor joint denoising, most on multi-hop WebQuestions. None of these probes is clean, so we state each caveat next to its result. Our contributions are:

- **Bidirectional Co-Refinement, stated precisely** as the commitment-order property of joint masked diffusion over a reasoning trace and a spoken reply, and contrasted with AR reasoning designs (Section 3.4).
- **DIFFUSPEECH**, a masked-diffusion LM for speech understanding and generation (ASR, TTS, S→T, S→S, T→T) and, to our knowledge, the first to denoise a reasoning trace jointly with a spoken reply, with the components needed beyond vocabulary extension identified (Section 3).
- **THINKINGTALK**, 26K spoken-QA samples (319 h) with reasoning traces, validated with cross-family judges and human raters, with its synthetic, post-hoc construction stated (Section 4).
- **A multi-probe study of the mechanism**, including a controlled 1.7B AR-vs.-diffusion comparison and matched-training ablations, with each probe's confound reported (Section 5).

2 RELATED WORK

Speech-text language models. SpiritLM (Nguyen et al., 2024) interleaves HuBERT units with text in one AR model. SpeechGPT (Zhang et al., 2023), Moshi (Défossez et al., 2024), Qwen2-Audio (Chu et al., 2024), GLM-4-Voice (Zeng et al., 2024), MinMo (Chen et al., 2025), Llama-Omni2 (Fang et al., 2025), MoST (Lou et al., 2026), Qwen2.5/3-Omni (Xu et al., 2025a;b), Step-Audio 2 (Wu et al., 2025a), and Audio Flamingo 3 (Goel et al., 2025) scale conversational and instruction-following ability. Ming-UniAudio (Yan et al., 2025) unifies understanding, generation, and editing in one LLM over continuous speech latents, with a diffusion head for synthesis. All of these produce their output sequence autoregressively.

Reasoning in spoken language models. TARS (Wang et al., 2026) narrows the gap between speech- and text-input reasoning, and Audio-Thinker (Wu et al., 2025d) learns when and how to think. For spoken output, STITCH, Mini-Omni-Reasoner, SHANKS, Mind-Paced Speaking (MPS; Wu et al., 2025c), and FLAIR (Wu et al., 2026), which reasons in a latent space while listening, overlap reasoning and speech in time, often at lower latency than DIFFUSPEECH. The property we

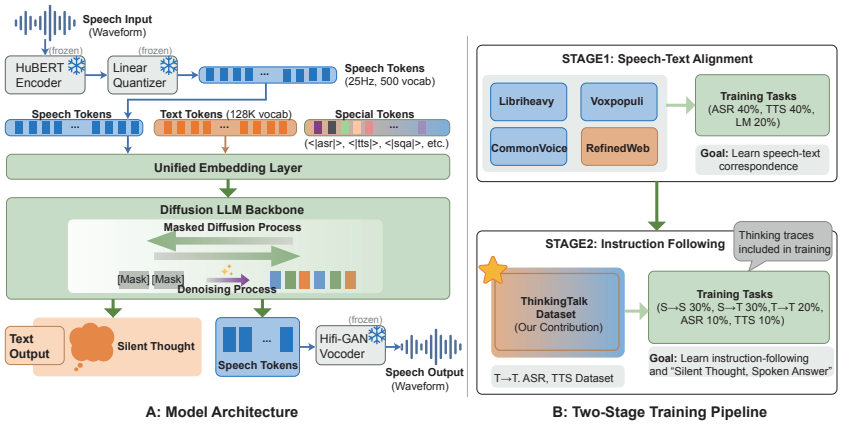

Figure 2: **DIFFUSPEECH. (a)** Frozen HuBERT units (25 Hz, K=500) share one vocabulary with text; a masked diffusion backbone initialized from LLaDA-8B denoises reasoning text and reply speech jointly; a frozen HiFi-GAN vocoder (Kong et al., 2020) synthesizes audio. **(b)** Stage 1 aligns speech and text (ASR, TTS, LM); Stage 2 adds S→S, S→T, and T→T with reasoning traces on THINKINGTALK.

study is different: whether a reasoning token can condition on speech that *follows* it in the output, which causal decoding rules out and joint masked diffusion permits. We do not claim that this makes diffusion better than these systems, which we have not compared under a shared protocol (Section 7).

Diffusion language models. Masked diffusion LMs (Sahoo et al., 2024; Shi et al., 2025) generate by iterative unmasking and now reach 7–8B parameters (Nie et al., 2025; Ye et al., 2025). Block diffusion (Arriola et al., 2025) and Fast-dLLM v2 (Wu et al., 2025b) decode blocks left to right to reuse a KV cache, and remasking samplers (Ghazvininejad et al., 2019; Wang et al., 2025b) can revise committed tokens. Multimodal variants handle images within one diffusion model (Yang et al., 2025b; You et al., 2025) or pair AR text with image diffusion (Transfusion; Zhou et al., 2025a). For speech, DiFFA (Zhou et al., 2025b) adapts LLaDA to speech understanding, diffusion LMs rescore ASR hypotheses (Naveriani et al., 2026), LLaDA-TTS (Fan et al., 2026) applies masked diffusion to TTS, and Omni-Diffusion (Li et al., 2026) covers ASR, TTS, and spoken visual QA, decoding a text response together with its spoken rendition. DIFFUSPEECH instead denoises a *reasoning trace*, whose content differs from the reply, jointly with the reply, uses modality-specific corruption, and studies how the two condition each other.

3 METHOD

3.1 UNIFIED SPEECH-TEXT REPRESENTATION

Speech and text share one vocabulary $\mathcal{V} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_{\text{special}} \cup \mathcal{V}_{\text{speech}}$ of 126,973 tokens (Appendix A); speech tokens are 25 Hz units from a frozen HuBERT encoder with a K=500 k-means quantizer, as in SpiritLM. A sample is a sequence $\boldsymbol{x} = [\tau, \boldsymbol{c}, \boldsymbol{y}]$ of a task token τ, a condition $\boldsymbol{c}$, and a target $\boldsymbol{y}$, with each text or speech segment wrapped in delimiter tokens. For speech-to-speech QA with reasoning (S2S),

$$\boldsymbol{x} = [\tau_{\text{s2s}},\ \underbrace{\boldsymbol{s}_{\text{user}}}_{\text{input speech}},\ \underbrace{\boldsymbol{t}_{\text{think}}}_{\text{reasoning (text)}},\ \underbrace{\boldsymbol{s}_{\text{reply}}}_{\text{reply speech}}],\qquad \boldsymbol{y} = [\boldsymbol{t}_{\text{think}}, \boldsymbol{s}_{\text{reply}}], \tag{1}$$

so the target holds both reasoning and reply positions (other formats: Appendix A.4).

3.2 MASKED DIFFUSION WITH MODALITY-SPECIFIC SCHEDULES

Given a clean sequence $\boldsymbol{x}_0 \in \mathcal{V}^n$ whose positions have known modalities $\mu_i \in \{\text{text}, \text{speech}\}$, the forward process masks each position independently at noise level $t \in [0, 1]$ with a modality-specific

probability,

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \prod_{i=1}^{n} \Big[\gamma_{\mu_i}(t)\, \mathbf{1}_{[x_t^{(i)} = \texttt{[M]}]} + \big(1 - \gamma_{\mu_i}(t)\big)\, \mathbf{1}_{[x_t^{(i)} = x_0^{(i)}]} \Big], \tag{2}$$

with $\gamma_{\text{text}}(t) = t$ and $\gamma_{\text{speech}}(t) = \cos\big(\frac{\pi}{2}(1-t)\big)$. A 40 ms speech unit is often predictable from its neighbors, so speech is masked more heavily than text at every noise level ($\gamma_{\text{speech}} \geq \gamma_{\text{text}}$ on $[0,1]$). MMaDA (Yang et al., 2025b) similarly decodes text and images with linear and cosine schedules at inference; we apply the distinction in the training-time corruption. A shared schedule hurts generation: with data fixed, LibriSpeech-clean TTS WER rises from 6.2 to 7.4 (shared cosine) or 8.1 (shared linear), and ASR WER from 3.0 to 3.5 or 3.4 (Table 10). Because both modalities share t, S2S training almost never presents a reply that is less masked than its trace, which matters in Section 5.1.

The reverse model f_θ is a transformer with bidirectional attention that predicts every masked position from all unmasked positions of both modalities, $p_\theta(x_0^{(i)} \mid \boldsymbol{x}_t) = \mathrm{softmax}(f_\theta(\boldsymbol{x}_t)_i)$. It is trained with the masked-diffusion cross-entropy, weighted by the inverse modality-specific schedule:

$$\mathcal{L}_{\text{MDLM}} = \mathbb{E}_{t, \boldsymbol{x}_t} \Big[\textstyle\sum_{i:\, x_t^{(i)} = \texttt{[M]}} -\gamma_{\mu_i}(t)^{-1} \log p_\theta(x_0^{(i)} \mid \boldsymbol{x}_t) \Big]. \tag{3}$$

3.3 Training

We adapt the two-stage alignment recipe of SpiritLM and Moshi to the diffusion objective (Figure 2b), always with *selective masking*: only target positions are corrupted and the condition stays clean, so the model learns $p_\theta(\boldsymbol{y} \mid \tau, \boldsymbol{c})$ (Equation (4)). **Stage 1** trains ASR, TTS, and text LM (proportions 0.4/0.4/0.2) on LibriHeavy (Kang et al., 2024), VoxPopuli, Common Voice, and RefinedWeb (Penedo et al., 2023). **Stage 2** trains S2S, S2T, and T2T with reasoning on ThinkingTalk while retaining ASR and TTS (0.3/0.3/0.2/0.1/0.1). In S2S samples, trace and reply are corrupted at the same noise level, so each modality learns to predict from partially decoded states of the other. Hyperparameters and a proportion-sensitivity analysis are in Appendices A.7 and D.11.

3.4 Inference and Bidirectional Co-Refinement

At inference, the target $\boldsymbol{y}$ starts as n mask tokens and is generated in one denoising process of T steps (Algorithm 1): at each step the model predicts every masked position, and the $\lceil n/T \rceil$ most confident positions, pooled across modalities, are committed. Masked positions are not assigned a modality in advance and no schedule is used: the model predicts the delimiters itself, and since text and speech occupy disjoint ID ranges, every committed token is unambiguously one or the other. Tokens after the predicted end delimiter are discarded; on the S→S benchmarks the median effective length is 731 tokens, 3.7% of generations reach $n_{\max}$=1024, and 0.6% have a malformed delimiter (Appendix A.6).

Algorithm 1 Joint denoising of trace and reply

Input: task τ, condition $\boldsymbol{c}$, length n, steps T
$\boldsymbol{x} \leftarrow [\tau, \boldsymbol{c}, \texttt{[M]}^{\,n}]$; $\Delta k \leftarrow \lceil n/T \rceil$
for $i = 1, \ldots, T$ **do**
  $p_j \leftarrow p_\theta(\cdot \mid \boldsymbol{x})$ for every masked j
  $\mathrm{conf}_j \leftarrow \max_v p_j(v)$
  $\mathcal{S} \leftarrow$ top-Δk masked j by conf_j
  $x^{(j)} \leftarrow \arg\max_v p_j(v)$ for $j \in \mathcal{S}$ ▷ *final*
end for
Split $\boldsymbol{y}$ at predicted delimiters into $\hat{\boldsymbol{t}}_{\text{think}}$, $\hat{\boldsymbol{s}}_{\text{reply}}$
return $\hat{\boldsymbol{t}}_{\text{think}}$, $\mathrm{Vocoder}(\hat{\boldsymbol{s}}_{\text{reply}})$

What co-refinement means here. Let π be the order in which target positions are committed. Under causal decoding π is the sequence order, so a reasoning position j is predicted from $\{x_k : k < j\}$ only. Under Algorithm 1, it is predicted at every step from all committed positions, including reply-speech positions $k > j$ committed earlier. *Bidirectional Co-Refinement* denotes exactly this two-way conditioning during joint denoising: committed reasoning conditions the remaining speech (the *forward* channel, which AR reasoning also has), and committed speech conditions the remaining reasoning, including reasoning that precedes it in the sequence (the *reverse* channel, which causal decoding cannot provide). Refinement acts on the predictive distributions of still-masked positions: none of 2.1M instrumented commitments changed afterwards, although a masked position's argmax can drift before commitment (Appendix D.10). Interleaved AR models let speech chunk k inform reasoning chunk $k{+}1$, never the reverse in sequence order; remasking samplers (Wang et al., 2025b), which would let speech revise committed reasoning, are left to future work.

From understanding to generation. Relative to an understanding-only diffusion LM such as DiFFA (Appendix D.12), DIFFUSPEECH needed an output-side speech vocabulary with a delimiter protocol (without which variable-length speech cannot be expressed), modality-specific corruption (without which TTS WER degrades by 1.2–1.9 points), and selective masking (not ablated separately). The fourth ingredient, joint denoising of trace and reply, is studied in Section 5.

4 THE THINKINGTALK DATASET

Joint denoising needs samples with a spoken question, a written reasoning trace, and a spoken answer; we know of no speech-QA resource with all three, so we build THINKINGTALK (26,387 samples, 319 h; Figure 3, Appendix B). **Rewriting.** Qwen3-32B (Yang et al., 2025a) rewrites conversational samples from SmolTalk2 (Hugging Face, 2025), the successor of SmolTalk (Allal et al., 2025), and logical-reasoning samples from SoundMind (Diao et al., 2025) into a spoken-style question, a step-by-step trace, and a speech-suitable reply without visual references or unverbalizable notation. **Filtering.** Qwen3-32B scores each sample on seven dimensions; samples with a weighted score below 3.5, or below 3 on correctness, oral suitability, or visual independence, are removed (26,387 of 32,292 kept). Re-scoring 2,000 samples with judges from other families agrees with it (GPT-4o (OpenAI, 2024): Spearman ρ=0.83, keep/reject κ=0.72; Llama-3.3-70B (Grattafiori et al., 2024): ρ=0.79), and on 150 samples rated by three humans, human–judge agreement (κ=0.69) approaches inter-human agreement (κ=0.74), with 93.3% of traces judged logically sound. **Synthesis.** Questions are voiced by MegaTTS3 (Jiang et al., 2025) with four LibriSpeech (Panayotov et al., 2015) reference speakers and replies by Qwen3-Omni in one fixed voice; a words-per-second check removes hallucinated audio. Samples average 43.5 s, with 77.5-word traces and 102-word replies. **Caveats.** All audio is synthetic and replies share one voice, so the answer side has little prosodic variety. Traces are written for a known answer and are thus post-hoc rationalizations, not records of the computation that produced it. Only 5.6% of samples come from the reasoning-oriented source; 88% are system-chat conversations, although every sample has a trace.

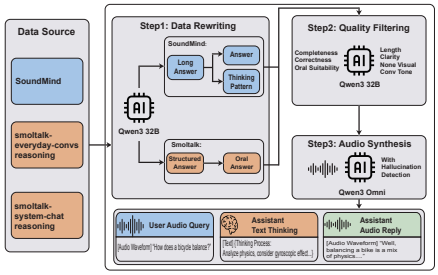

Figure 3: THINKINGTALK construction: rewrite text QA into spoken style with an explicit trace, filter with an LLM judge, and synthesize question and reply audio.

5 PROBING BIDIRECTIONAL CO-REFINEMENT

If joint denoising helped only because the reply reads the trace, a think-then-speak AR model would share the benefit. The question is whether the *reverse* channel—speech conditioning still-masked reasoning—contributes. No single experiment isolates it cleanly, so we combine probes with different confounds: inference-time interventions at 8B (Section 5.1), matched-training ablations and a controlled AR-vs.-diffusion comparison at 1.7B (Section 5.2), error injection (Section 5.3), and decoding instrumentation (Section 5.4). S→S accuracy transcribes the spoken reply with Whisper-large-v3 (Radford et al., 2022); decoding uses T=512 steps unless stated; protocols are in Appendix C.

5.1 WHICH DIRECTION CARRIES THE BENEFIT?

Table 1 changes only how the 8B model decodes, with the same weights and 512-step budget. **Forward channel.** Blocking the reply from attending to the trace (44.6) or denoising the reply before the trace (44.8) leaves no gain over omitting the trace (44.4), so the benefit of reasoning reaches the answer through the trace. Freezing a manipulated trace before the reply confirms that its content is used: accuracy falls from 47.8 (own trace) to 45.6 (truncated to 25%), 42.3 (shuffled), and 41.0 (irrelevant), the last two below having no trace (Appendix C.2). **Reverse channel.** Freezing the fully denoised trace before any speech is decoded—the AR order inside our own weights—costs 2.8 points, and preventing reasoning from attending to speech costs 2.7. Both losses concentrate on WebQuestions (−4.8 and −4.5), the most multi-hop benchmark, and are smallest on LlamaQuestions

Table 1: **Inference-time interventions on the 8B model** (S→S accuracy, %; same weights; 512 steps). The forward channel is reasoning→speech, the reverse channel speech→reasoning. Δ: change from joint denoising.

Decoding regime	LlamaQ	TriviaQA	WebQ	Avg	Δ
Joint denoising (default)	68.5	33.5	49.7	50.6	–
Reverse channel removed					
Reasoning cannot attend to speech	66.6	31.9	45.2	47.9	−2.7
Think-first: trace fully denoised and frozen before speech	66.8	31.6	44.9	47.8	−2.8
Forward channel removed					
Speech cannot attend to reasoning	65.1	28.0	40.8	44.6	−6.0
Speech-first: reply denoised before the trace	65.0	28.4	41.0	44.8	−5.8
Both removed					
Both directions blocked	64.7	27.6	40.1	44.1	−6.5
No reasoning segment	65.2	27.8	40.3	44.4	−6.2

Table 2: **Controlled experiments at 1.7B** (Qwen3-1.7B architecture, random init.; S→S accuracy, %). **(a)** AR vs. diffusion variants differing only in the attention mask. **(b)** Diffusion Stage-2 arms differing only in the corruption pattern, each decoded in its trained regime. Every row is a single training run (caveat in Section 5.2).

		LlamaQ	TriviaQA	WebQ	Avg
(a) AR vs. diffusion × THINKINGTALK					
AR (causal)	w/o THINKINGTALK	–	–	–	17.8
AR (causal)	w/ THINKINGTALK	38.5	12.8	18.4	23.2
Diffusion (bidirectional)	w/o THINKINGTALK	–	–	–	18.5
Diffusion (bidirectional)	w/ THINKINGTALK	42.3	15.1	22.7	26.7
(b) Matched-training arms (diffusion, w/ THINKINGTALK; separate runs from (a))					
A: joint masking	joint decoding	42.5	17.3	27.6	29.1
B: think-first curriculum	think-first decoding	41.1	16.6	22.8	26.8
C: reasoning attends to text only	same at inference	38.7	16.3	21.4	25.5

(−1.7 and −1.9). **Caveat.** These interventions put the model in states it was not trained on: with a shared noise level and $\gamma_{\text{speech}} \geq \gamma_{\text{text}}$, S2S training essentially never shows a finished trace beside fully masked reply speech (TTS samples do show clean text conditioning masked speech), and the attention masks never occur in training. Part of each loss may reflect distribution shift rather than a severed channel; Section 5.2 removes this confound at a smaller scale.

5.2 MATCHED TRAINING AND A CONTROLLED COMPARISON AT 1.7B

Controlled AR vs. diffusion. We train two variants of the Qwen3-1.7B dense architecture (Yang et al., 2025a) from random initialization with identical tokenizer, vocabulary, speech units, data, hyperparameters, and curriculum; *only* the attention mask differs. With THINKINGTALK, the diffusion variant leads on every benchmark (26.7 vs. 23.2 S→S average) and on LibriSpeech ASR (3.8 vs. 4.2 WER) and TTS (7.8 vs. 8.5); without it, the two are close (18.5 vs. 17.8), so the gap appears mainly when traces are available (+8.2 vs. +5.4; Table 2a). This controls for scale, data, and initialization, but not for *how* the trace is used: bidirectional attention might simply model the trace better, independent of speech-to-reasoning conditioning. **Matched-training arms.** To separate the reverse channel from that explanation and from the distribution shift above, we train three Stage-2 arms of the 1.7B diffusion variant with identical data, steps, and hyperparameters, differing only in the corruption pattern, and decode each in its trained regime with $T{=}512$ steps. *Arm A* uses joint masking and joint decoding. *Arm B* is a think-first curriculum—reasoning is corrupted only while the reply is fully masked, the reply only while the trace is clean—decoded think-first, i.e., the sequential regime made in-distribution. *Arm C* restricts reasoning positions to attend only to text, in training and at inference. Joint denoising leads on every benchmark (A 29.1, B 26.8, C 25.5), with the largest gaps on WebQuestions (27.6 vs. 22.8 and 21.4); even Arm B, trained for exactly the regime it is decoded in, trails. **Caveat.** Arm A nominally matches the diffusion-with-THINKINGTALK row of Table 2a but was trained separately and scores 2.4 points higher (29.1 vs. 26.7, mostly on WebQuestions). With

one run per configuration, we cannot separate run-to-run variation from an unintended configuration difference, and a difference this large matches the A–B gap (2.3). We therefore claim only the ordering A > B > C, which holds on all three benchmarks, not the size of the gaps, which requires multi-seed runs.

5.3 Error Injection

Error injection asks whether the reverse channel helps correct reasoning. For 300 examples (100 per benchmark) that the 8B model answers correctly, we force-commit, after 25% of the steps, one reasoning token that corrupts the key entity or number. Under default decoding the spoken answer stays correct in 64% of runs. The corruption propagates—mean KL over still-masked reasoning positions rises to 0.19 (0.031 after an ordinary commitment), and 71% of injections flip a downstream reasoning argmax within three steps—and in 57% of recovered runs the later trace explicitly overrides the corrupted value. When reasoning cannot attend to speech, recovery falls to 41%. The 64% is on-trajectory, but the 41% control shares the distribution-shift caveat of Section 5.1; the more direct control, the same injection under think-first decoding, was not run.

5.4 Decoding Instrumentation

Logging Algorithm 1 on about 200 examples per benchmark (Table 3) shows that the reverse channel is available and used. 41% of reasoning tokens are committed after the first speech token (46% on WebQuestions), and the share of reasoning-position attention on committed speech grows from 2.9% in the first quarter of steps to 18.4% in the last. Re-running the forward pass with a just-committed speech token re-masked, 17.4% of speech commitments flip at least one still-masked reasoning argmax (mean KL 0.031, top-decile mean 0.24). These statistics rise from LlamaQuestions to WebQuestions, as do the gain from reasoning and the cost of freezing the trace (Table 1). They are descriptive: there is no null condition, and re-masking *any* committed token, reasoning included, would perturb some argmaxes, so the flip rate cannot by itself be attributed to the speech content.

Table 3: **Decoding statistics** (8B, ∼200 examples per benchmark). The first three rows have no null condition.

	LlamaQ	TriviaQA	WebQ
Reasoning after 1st speech (%)	38	40	46
Speech commits w/ flip (%)	14.9	16.9	20.4
Mean KL per speech commit	.026	.031	.039
Gain from reasoning	3.3	5.7	9.4
Cost of think-first	1.7	1.9	4.8

What the probes support. The forward channel accounts for most of the value of reasoning, as in AR chain-of-thought. The reverse channel, specific to joint denoising, is used during decoding, and every intervention that removes it lowers accuracy, most on multi-hop questions: by 2.3–3.6 points under matched training at 1.7B and by 2.7–2.8 at inference time at 8B. Each probe leaves a gap (single runs, off-distribution states, a non-native control, no null condition), and the probes do not measure one common quantity. We read them as consistent support for a modest reverse-channel contribution, not as a demonstration of it.

6 System-Level Evaluation

6.1 Setup

Baselines and benchmarks. We compare with speech-generative AR models up to 10B parameters (SpeechGPT, SpiritLM, Moshi, MinMo, Llama-Omni2, Qwen2.5-Omni, GLM-4-Voice, Step-Audio 2 mini), S→T-only models (Qwen2-Audio, Phi-4-Multimodal (Microsoft et al., 2025), Audio Flamingo 3, DiFFA), and the base LLaDA. These AR baselines are general-purpose speech LMs, not reasoning-while-speaking systems; we did not run STITCH or MPS under our protocol. Benchmarks are spoken QA—LlamaQuestions (Nachmani et al., 2024), TriviaQA (Joshi et al., 2017), and WebQuestions (Berant et al., 2013) in S→T and S→S modes, OpenBookQA (Mihaylov et al., 2018) in S→T mode—ASR/TTS on LibriSpeech, VoxPopuli (Wang et al., 2021), and Common Voice (Ardila et al., 2020), text benchmarks (MMLU (Hendrycks et al., 2021a), TriviaQA, GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), HumanEval (Chen et al., 2021)), and MMSU (Wang et al., 2025a) (Appendix D.5). **Evaluation protocol.** S→S accuracy transcribes the reply with Whisper-large-v3 and applies standard answer normalization; TTS WER compares the transcript with the

Table 4: **Spoken QA** (accuracy, %). Averages and $\Delta_{T\to S}$ (S→T minus S→S) are over LlamaQ, TriviaQA, and WebQ, the benchmarks with both modes; OBQA is S→T only. **Bold**: best; underline: second. '/': mode not supported; '–': not evaluated. Evaluation protocol: Section 6.1 and Appendix D.1.

				LlamaQ		TriviaQA		WebQ		OBQA	Average		
Model	**Type**	**Params**	**S→S**	S→T	S→S	S→T	S→S	S→T	S→S	S→T	S→T	S→S	$\Delta_{T\to S}$
Qwen2-Audio	AR	8B	✗	55.4	/	38.5	/	41.2	/	49.5	45.0	/	/
Phi-4-Multimodal	AR	5.6B	✗	60.2	/	22.8	/	26.6	/	**65.9**	36.5	/	/
Audio Flamingo 3	AR	7B	✗	76.4	/	44.2	/	58.8	/	–	59.8	/	/
SpeechGPT	AR	7B	✓	21.6	5.4	14.8	8.2	6.3	5.5	12.4	14.2	6.4	7.9
SpiritLM	AR	2.7B	✓	40.1	38.7	6.5	4.2	12.3	9.6	21.7	19.6	17.5	2.1
Moshi	AR	7B	✓	68.3	64.2	42.1	30.5	47.5	40.7	26.0	52.6	45.1	7.5
MinMo	AR	8B	✓	**76.5**	63.8	40.1	25.5	55.0	39.9	44.5	57.2	43.1	14.1
Llama-Omni2	AR	8B	✓	70.3	60.7	35.3	23.9	40.4	37.1	28.1	48.7	40.6	8.1
Qwen2.5-Omni	AR	7B	✓	73.8	65.5	42.6	29.8	53.2	42.1	55.8	56.5	45.8	10.7
GLM-4-Voice	AR	9B	✓	64.7	50.7	39.1	26.5	32.2	15.9	36.0	45.3	31.0	14.3
Step-Audio 2 mini	AR	8B	✓	75.2	67.6	**47.8**	**36.2**	**62.6**	**52.8**	–	**61.9**	**52.2**	9.7
DiFFA	Diff.	8B	✗	58.3	/	36.0	/	43.4	/	35.6	45.9	/	/
DIFFUSPEECH	Diff.	8B	✓	72.0	**68.5**	45.4	33.5	61.5	49.7	51.3	59.6	50.6	9.1

Table 5: **(a) Text abilities** (T→T) of DIFFUSPEECH and its base model. **(b) ASR/TTS WER** (% ↓, LibriSpeech clean/other), **UTMOS** (↑, clean), and within-response **speaker similarity** (↑). †: copied from the cited source. Full results: Appendices D.2 to D.4.

(a)	**LLaDA**	**DIFFUSPEECH**
MMLU	65.9	66.2 +0.3
TriviaQA	55.6	60.3 +4.7
GSM8K	70.3	72.8 +2.5
MATH	32.5	33.8 +1.3
HumanEval	38.4	40.1 +1.7

(b)	**ASR**	**TTS**	**UTMOS**	**SpkSim**
Qwen2.5-Omni	1.8 / 3.6	5.0 / 5.4	4.14	0.958
Step-Audio 2 mini	1.33† / 2.86†	4.3 / 5.5	–	–
GLM-4-Voice	2.8 / 4.0	5.8 / 6.3	4.02	0.941
MinMo	1.8 / 3.9	6.7 / 7.5	3.97	0.946
Moshi	5.5 / 12.0	7.0 / 7.2	3.71	0.928
DIFFUSPEECH	3.0 / 4.3	6.2 / 6.8	3.86	0.947
Ground truth	–	–	4.09	–

input text; cells marked † are copied from the cited source. Absolute numbers depend on the protocol: GLM-4-Voice's LlamaQuestions (S→T) accuracy is 64.7 in its own report (Zeng et al., 2024) but 74.3 under the protocol of STITCH (Chiang et al., 2026). DIFFUSPEECH decodes with $T{=}512$ and $n_{\max}{=}1024$, without acceleration, unless noted.

6.2 SPOKEN QUESTION ANSWERING

On S→S, DIFFUSPEECH averages 50.6% (Table 4; 50.6 ± 0.3 over five inference seeds), ahead of Qwen2.5-Omni (45.8), Moshi (45.1), MinMo (43.1), and GLM-4-Voice (31.0), and behind Step-Audio 2 mini (52.2). It is best on LlamaQuestions S→S and second on the other two. On S→T it is third on average (59.6, vs. 61.9 for Step-Audio 2 mini and 59.8 for Audio Flamingo 3) and 13.7 points above the other diffusion LM, DiFFA. **S→T-to-S→S degradation.** DIFFUSPEECH loses 9.1 points when the answer must be spoken (bootstrap 95% CI [8.3, 9.9]), less than Qwen2.5-Omni (10.7), MinMo (14.1), and GLM-4-Voice (14.3), but not distinguishably less than Step-Audio 2 mini (9.7); weaker systems (e.g., Moshi 7.5, SpiritLM 2.1) lose less from a lower base. The advantage is uneven: it comes from LlamaQuestions (3.5 vs. 7.6 for Step-Audio 2 mini), while Step-Audio 2 mini preserves more on TriviaQA and WebQuestions (Appendix D.6). A small drop is consistent with joint denoising but is correlational, with other plausible causes (speech-path fidelity, tokenization loss, how well Whisper transcribes each system), so we do not count it as evidence for the mechanism.

6.3 Text Abilities and Speech Quality

Text. Speech-text training does not erode the backbone: DiffuSpeech matches or slightly exceeds LLaDA on all five text benchmarks (Table 5a); otherwise its text ability is inherited, and AR models with stronger base LLMs score higher (Appendix D.4). **Speech.** ASR and TTS are competitive but not state-of-the-art (Table 5b): Qwen2.5-Omni and Step-Audio 2 mini, trained on far larger speech corpora, transcribe and synthesize more accurately. DiffuSpeech's UTMOS (Saeki et al., 2022) (3.86) lies between Moshi (3.71) and MinMo (3.97), below Qwen2.5-Omni (4.14), and its within-response speaker similarity (WavLM embeddings; Chen et al., 2022) matches the AR systems. The naturalness gap is consistent with the acoustic ceiling of discrete HuBERT units (Nguyen et al., 2024); we have no subjective MOS or prosody evaluation.

6.4 Operating Point and Cost

Table 6 reports accuracy and latency at the same operating points. Vanilla decoding is slower than real time (RTF 1.85). The training-free Fast-dLLM v2 accelerator (block-wise decoding with a KV cache) cuts total time to 6.2 s (RTF 0.62) for a loss of 0.9 points, mostly on WebQuestions (-1.3), and curtails the reverse channel (flip rate 17.4% $\rightarrow$ 12.1%). Time to first audio (1.0 s) remains well above streaming AR (0.2 s for Moshi), so the system suits turn-based rather than full-duplex dialogue. With T fixed, long-form TTS WER rises from 6.2 at n=512 to 9.6 at n=2048; T=n recovers most of it (7.2) but makes wall-clock grow linearly in n, as in AR decoding (Appendix D.8). AR latencies from the same unoptimized harness are in Appendix D.7 as context, not as a ranking. An ASR$\rightarrow$LLaDA$\rightarrow$TTS cascade reaches 46.7% at 8.5 s, below the accelerated DiffuSpeech (49.7% at 6.2 s) on every benchmark (Appendix D.9).

Table 6: **Accuracy and latency at one operating point** (S$\rightarrow$S Avg, %; one H100, batch 1, bf16; $\sim$5 s question, $\sim$10 s reply). Flip: % of speech commitments that flip $\geq$1 reasoning argmax.

Setting	Avg	TTFA	Total	RTF	Flip
Vanilla, T=512	50.6	3.2 s	18.5 s	1.85	17.4
+ Fast-dLLM v2	49.7	1.0 s	6.2 s	0.62	12.1

7 Limitations and Conclusion

Limitations. **Mechanism.** The reverse-channel evidence rests on imperfect probes: inference-time interventions leave the training distribution, the matched-training arms are single runs and a separate run of the joint configuration differs by 2.4 points, the error-injection control is off-distribution, and the decoding statistics lack a null condition. **Comparisons.** We did not evaluate reasoning-while-speaking AR systems (STITCH, MPS) under our protocol; our AR comparisons are general-purpose models and a think-then-speak twin that cannot condition reasoning on speech by construction. **Latency.** Time to first audio exceeds streaming AR and grows with the reasoning prefix under block decoding; at matched long-form quality, wall-clock grows linearly with length. **Data and speech.** ThinkingTalk is synthetic, its traces are post-hoc, its reasoning-oriented share is small, and replies use one voice. Discrete HuBERT units cap naturalness; a hybrid of discrete text and continuous audio latents, as in Transfusion (Zhou et al., 2025a), is a natural next step. We have no subjective MOS or prosody evaluation, train each 8B configuration once, and cover only English short-answer QA.

Conclusion. DiffuSpeech shows that one masked diffusion LM can cover speech understanding and generation while denoising a reasoning trace jointly with a spoken reply, at accuracy competitive with AR speech LMs of its size. Most of the benefit of reasoning flows from trace to speech; the reverse channel that joint denoising adds is used, and removing it consistently lowers accuracy, most on multi-hop questions, but its size remains to be established with multi-seed matched training and null-conditioned measurements.

Ethics Statement

DiffuSpeech generates speech and could be misused for impersonation. It speaks in a single fixed assistant voice and is trained only on publicly available data. ThinkingTalk is derived from SmolTalk2 (Apache 2.0) and SoundMind (MIT), with question audio synthesized from LibriSpeech reference voices (CC BY 4.0); we will release it under compatible terms. Reasoning traces make

outputs easier to inspect, but because the model is trained on post-hoc rationalizations, its traces should not be read as faithful accounts of its computation.

REPRODUCIBILITY STATEMENT

Appendix A specifies the vocabulary, sequence formats, masking schedules, selective-masking objective, training and inference hyperparameters, and hardware. Appendix B gives the prompts, judge weights, filtering thresholds, and synthesis settings used to build THINKINGTALK. Appendix C describes each mechanism probe, and Appendix D.1 documents the evaluation harness. The base model (LLaDA-8B-Instruct), speech tokenizer (HuBERT with 500-unit k-means, as in SpiritLM), vocoder (HiFi-GAN), and all baselines are publicly available. We will release code, checkpoints, and THINKINGTALK upon acceptance. Each configuration was trained once (the 8B model takes 22 days on 32 H20 GPUs for Stage 1); we report inference-seed variance, not training-seed variance.

AI USE STATEMENT

We used generative AI in two ways. (i) As part of the method, Qwen3-32B rewrote and scored THINKINGTALK samples, GPT-4o and Llama-3.3-70B re-scored a subset for validation, and Qwen3-Omni and MegaTTS3 synthesized its audio (Section 4). (ii) LLM-based assistants helped edit and restructure the manuscript text and convert it to this template. Generative AI was not used to produce experimental results. The authors reviewed all AI-assisted text, checked every reported number against their experiment logs, and take full responsibility for the content of this paper.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, et al. SmolLM2: When smol goes big – data-centric training of a small language model. arXiv preprint arXiv:2502.02737, 2025.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670, 2020.

Marianne Arriola, Aaron Gokaslan, Justin T. Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, 2025.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1160.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, et al. MinMo: A multimodal large language model for seamless voice interaction. arXiv preprint arXiv:2501.06282, 2025.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022. doi: 10.1109/JSTSP.2022.3188113.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. arXiv preprint arXiv:2410.17196, 2024.

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, 2025.

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, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. arXiv preprint arXiv:2407.10759, 2024.

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.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. SoundMind: RL-incentivized logic reasoning for audio-language models. arXiv preprint arXiv:2506.12935, 2025.

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

Xiaoyu Fan, Huizhi Xie, Wei Zou, and Yunzhang Chen. LLaDA-TTS: Unifying speech synthesis and zero-shot editing via masked diffusion modeling. arXiv preprint arXiv:2603.26364, 2026.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. LLaMA-Omni2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. arXiv preprint arXiv:2505.02625, 2025.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. doi: 10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. arXiv preprint arXiv:2103.03874, 2021b.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. arXiv preprint arXiv:2106.07447, 2021.

Hugging Face. SmolTalk2: Post-training data for SmolLM3. `https://huggingface.co/datasets/HuggingFaceTB/smoltalk2`, 2025.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. MegaTTS 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. arXiv preprint arXiv:2502.18924, 2025.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551, 2017.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: a 50,000 hours ASR corpus with punctuation casing and context. arXiv preprint arXiv:2309.08105, 2024.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint arXiv:2010.05646, 2020.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-Diffusion: Unified multimodal understanding and generation with masked discrete diffusion. In *International Conference on Machine Learning*, 2026.

Yuxuan Lou, Kai Yang, and Yang You. MoST: Mixing speech and text with modality-aware mixture of experts. arXiv preprint arXiv:2601.10272, 2026.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv preprint arXiv:2503.01743, 2025.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. arXiv preprint arXiv:1809.02789, 2018.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. arXiv preprint arXiv:2305.15255, 2024.

Davyd Naveriani, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Diffusion language models for speech recognition. arXiv preprint arXiv:2604.14001, 2026.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. Spirit LM: Interleaved spoken and written language model. arXiv preprint arXiv:2402.05755, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. arXiv preprint arXiv:2502.09992, 2025.

OpenAI. GPT-4o system card. arXiv preprint arXiv:2410.21276, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356, 2022.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proc. Interspeech*, 2022.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. arXiv preprint arXiv:2406.07524, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. arXiv preprint arXiv:2406.04329, 2025.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390, 2021.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. arXiv preprint arXiv:2601.05543, 2026.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. arXiv preprint arXiv:2506.04779, 2025a.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903, 2023.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, et al. Step-Audio 2 technical report. arXiv preprint arXiv:2507.16632, 2025a.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. Fast-dLLM v2: Efficient block-diffusion LLM. arXiv preprint arXiv:2509.26328, 2025b.

Donghang Wu, Haoyang Zhang, Jun Chen, Xiangyu (Tony) 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. arXiv preprint arXiv:2510.09592, 2025c.

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. arXiv preprint arXiv:2603.17837, 2026.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-Thinker: Guiding audio language model when and how to think via reinforcement learning. arXiv preprint arXiv:2508.08039, 2025d.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Mini-Omni-Reasoner: Token-level thinking-in-speaking in large speech models. arXiv preprint arXiv:2508.15827, 2025.

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. arXiv preprint arXiv:2503.20215, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-Omni technical report. arXiv preprint arXiv:2509.17765, 2025b.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, et al. Ming-UniAudio: Speech LLM for joint understanding, generation and editing with unified representation. arXiv preprint arXiv:2511.05516, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models. arXiv preprint arXiv:2508.15487, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933, 2025.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025a.

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. arXiv preprint arXiv:2507.18452, 2025b.

A Implementation Details

A.1 Notation

Table 7: Notation used in the paper.

Symbol	Meaning
$\mathcal{V} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_{\text{special}} \cup \mathcal{V}_{\text{speech}}$	Unified vocabulary and its parts
$\boldsymbol{x} = [\tau, \boldsymbol{c}, \boldsymbol{y}]$	Sequence: task token, condition, target
$\boldsymbol{s}_{\text{user}}, \boldsymbol{t}_{\text{think}}, \boldsymbol{s}_{\text{reply}}$	User speech, reasoning trace (text), reply speech
$\mathcal{C}, \mathcal{Y}$	Condition and target positions
$\mu_i \in \{\text{text}, \text{speech}\}$	Modality of position i
t; $\gamma_{\text{text}}(t), \gamma_{\text{speech}}(t)$	Noise level; per-modality masking probabilities
$\boldsymbol{x}_t$, `[M]`	Corrupted sequence at noise level t; mask token
f_θ, p_θ	Denoiser and its predictive distribution
T, n, $\Delta k = \lceil n/T \rceil$	Denoising steps, target length, commitments per step
π	Order in which target positions are committed
Forward / reverse channel	Reasoning→speech / speech→reasoning conditioning during joint denoising
$\Delta_{T \to S}$	S→T accuracy minus S→S accuracy
TTFA, RTF	Time to first audio; real-time factor (generation time / audio duration)

A.2 Vocabulary and Special Tokens

Table 8: Vocabulary of DiffuSpeech (top) and special tokens (bottom).

Component	Size	Token IDs
Text (from LLaDA)	126,464	0–126,463
Special	9	126,464–126,472
Speech (HuBERT units, K=500)	500	126,473–126,972
Total	**126,973**	—

Tokens	Function
`<|t2s|>`, `<|asr|>`, `<|s2s|>`, `<|s2t|>`, `<|t2t|>`	Task selectors
`<|sot|>`, `<|eot|>`	Text segment delimiters
`<|sos|>`, `<|eos|>`	Speech segment delimiters

A.3 Speech Tokenization

We follow the SpiritLM (Nguyen et al., 2024) pipeline: a HuBERT-base encoder pretrained on 960 h of LibriSpeech; a k-means quantizer with K=500 on HuBERT features; a frame rate of 25 Hz (one unit per 40 ms); and a HiFi-GAN vocoder (Kong et al., 2020) that reconstructs audio from units. All speech components are frozen during DiffuSpeech training.

A.4 Sequence Formats

Table 9: Sequence formats. Only target positions are corrupted in training and generated at inference; all other positions are clean conditions.

Task	Sequence
TTS	`<|t2s|>` `<|sot|>` $\boldsymbol{t}_{\text{in}}$ `<|eot|>` `<|sos|>` $\boldsymbol{s}_{\text{out}}$ `<|eos|>` target: $\boldsymbol{s}_{\text{out}}$
ASR	`<|asr|>` `<|sos|>` $\boldsymbol{s}_{\text{in}}$ `<|eos|>` `<|sot|>` $\boldsymbol{t}_{\text{out}}$ `<|eot|>` target: $\boldsymbol{t}_{\text{out}}$
S2S	`<|s2s|>` `<|sos|>` $\boldsymbol{s}_{\text{user}}$ `<|eos|>` `<|sot|>` $\boldsymbol{t}_{\text{think}}$ `<|eot|>` `<|sos|>` $\boldsymbol{s}_{\text{reply}}$ `<|eos|>` target: $\boldsymbol{t}_{\text{think}}$ and $\boldsymbol{s}_{\text{reply}}$, denoised jointly
S2T	`<|s2t|>` `<|sos|>` $\boldsymbol{s}_{\text{user}}$ `<|eos|>` `<|sot|>` $\boldsymbol{t}_{\text{think}}$ $\boldsymbol{t}_{\text{reply}}$ `<|eot|>` target: $\boldsymbol{t}_{\text{think}}, \boldsymbol{t}_{\text{reply}}$
T2T	`<|t2t|>` `<|sot|>` $\boldsymbol{t}_{\text{user}}$ `<|eot|>` `<|sot|>` $\boldsymbol{t}_{\text{think}}$ $\boldsymbol{t}_{\text{reply}}$ `<|eot|>` target: $\boldsymbol{t}_{\text{think}}, \boldsymbol{t}_{\text{reply}}$

A.5 MODALITY-SPECIFIC MASKING

Every position i is tagged with its modality μ_i at data-loading time (segments are delimited by special tokens). At each training step we draw one noise level $t \sim U(0,1)$ per sample and mask each position independently with probability $\gamma_{\mu_i}(t)$. Both schedules use the same t, so within a sample text and speech are masked at consistent but different rates; because $\cos(\frac{\pi}{2}(1-t)) = \sin(\frac{\pi t}{2}) \geq t$ on $[0,1]$, speech is never less masked than text in expectation. Table 10 ablates the choice.

Table 10: Effect of modality-specific masking schedules (DIFFUSPEECH after Stage 1; LibriSpeech-clean WER and Wikitext perplexity).

Schedule	ASR WER	TTS WER	Wikitext PPL
Cosine (both modalities)	3.5	7.4	6.51
Linear (both modalities)	3.4	8.1	6.32
Linear text + cosine speech (DIFFUSPEECH)	**3.0**	**6.2**	**6.30**

A.6 TARGET LENGTH, DELIMITERS, AND MODALITY BOUNDARIES

No separate length predictor is used. Each task has a maximum target length $n_{\max}$ (1024 for S2S, S2T, and T2T; 512 for ASR and TTS). The target is initialized with $n_{\max}$ mask tokens, the model is trained to emit end delimiters, and tokens after the first predicted end delimiter are discarded. Padding positions are processed in the same forward passes, so this adds no latency beyond what we report. On the three S→S benchmarks, the effective target length has a median of 731 tokens (5th–95th percentile 402–988); 3.7% of generations reach $n_{\max}$ and are truncated, and 0.6% have a missing or malformed delimiter, in which case we truncate at $n_{\max}$.

Modality boundaries at inference. The schedules of Equation (2) are used only in the training-time corruption process, where each position's modality is known from the clean sample. Inference (Algorithm 1) uses no schedule and does not assign modalities to masked positions in advance: segment boundaries arise from the delimiter tokens the model predicts, and because text and speech occupy disjoint ID ranges (Table 8), every committed token is unambiguously text or speech.

A.7 TRAINING CONFIGURATION

Selective masking and multi-task objective. For a sample $\boldsymbol{x} = [\tau, \boldsymbol{c}, \boldsymbol{y}]$ of task k with condition positions $\mathcal{C}$ and target positions $\mathcal{Y}$, only the target is corrupted:

$$x_t^{(i)} = \begin{cases} x_0^{(i)} & \text{if } i \in \mathcal{C}, \\ \texttt{[M]} \text{ w.p. } \gamma_{\mu_i}(t), \text{ else } x_0^{(i)} & \text{if } i \in \mathcal{Y}, \end{cases} \tag{4}$$

so the model learns $p(\boldsymbol{y} \mid \boldsymbol{c}, \tau)$. With per-task sampling probabilities p_k and loss coefficients λ_k, the objective is

$$\mathcal{L} \;=\; \sum_{k \in \mathcal{T}} \lambda_k \cdot \mathbb{E}_{\boldsymbol{x}^{(k)}, t} \left[\sum_{i \in \mathcal{Y}_k :\, x_t^{(i)} = \texttt{[M]}} -\frac{1}{\gamma_{\mu_i}(t)} \log p_\theta(x_0^{(i)} \mid \boldsymbol{x}_t^{(k)}) \right]. \tag{5}$$

Table 11: Training configuration of the two stages (8B model).

Parameter	Stage 1 (alignment)	Stage 2 (instruction following)
Initialization	LLaDA-8B-Instruct	Stage-1 model
Tasks	TTS, ASR, LM	S2S, S2T, T2T, ASR, TTS
Task proportions p_k	0.4, 0.4, 0.2	0.30, 0.30, 0.20, 0.10, 0.10
Loss coefficients λ_k	1.0 (all)	1.0 (all)
Data	LibriHeavy (29K h), VoxPopuli, Common Voice; RefinedWeb (text)	THINKINGTALK (26,387 samples), Common Voice, VoxPopuli
Sequence length	1024	2048
Batch size	32 per GPU	–
Hardware	32 × NVIDIA H20, 22 days	8 × NVIDIA H20
Optimizer	AdamW, lr 1×10^{-5}, cosine decay, 1000-step warmup	AdamW, lr 5×10^{-6}

Controlled 1.7B variants. The AR and diffusion variants of Section 5.2 use the Qwen3-1.7B dense architecture from random initialization, the same vocabulary and speech units as DIFFUSPEECH, and the same two-stage data and hyperparameters; they differ only in the attention mask (and hence in the objective: next-token prediction vs. Equation (5)).

A.8 INFERENCE CONFIGURATION

All main results use T=512 denoising steps, confidence-based unmasking with $\Delta k = \lceil n/T \rceil$ positions per step pooled across modalities, temperature 1.0, and $n_{\max}$ of 1024 (S2S, S2T, T2T) or 512 (ASR, TTS).

B THINKINGTALK DETAILS

B.1 LLM PROMPTS

Stage-1 rewriting prompt (Qwen3-32B)

```
You are rewriting a text QA pair into an oral-style spoken QA
with explicit reasoning.  Given the input QA below, produce three
components:
(1) user_input:  a short, natural spoken question, no visual references
/ formulas / code.
(2) assistant_thinking:  a step-by-step reasoning trace (60-100 words)
that explains how to answer.
(3) assistant_reply:  a friendly spoken answer (80-120 words) that
reflects the reasoning above.  Avoid bullet points.
Input QA: {{qa_pair}}.
Output JSON: {"user_input":  ..., "assistant_thinking":  ...,
"assistant_reply":  ...}.
```

Stage-2 quality-judge prompt (Qwen3-32B)

```
Score the candidate sample on each of the seven dimensions below from 1
(worst) to 5 (best).  Provide a brief justification per dimension.
Dimensions:  Completeness, Correctness, Oral suitability, Clarity,
Length appropriateness, Visual independence, Conversational tone.
Sample:  {{user_input, assistant_thinking, assistant_reply}}.
Output JSON with all seven scores and an overall reasoning.
```

B.2 JUDGE CRITERIA AND VALIDATION

Table 12: Quality-judge dimensions and weights. A sample is kept if and only if its weighted score is at least 3.5 and each of correctness, oral suitability, and visual independence is at least 3.

Dimension	Description	Weight
Completeness	Response fully addresses the question	1.0
Correctness	Information is accurate and logical	1.5
Oral suitability	Appropriate for spoken delivery	1.5
Clarity	Clear and well structured	1.0
Length appropriateness	Suitable length for speech	0.8
Visual independence	No reliance on visual elements	1.2
Conversational tone	Natural when spoken	1.0

Validation. The weights are hand-set. To test whether the Qwen3-32B judge is lenient toward Qwen-family text, we re-scored 2,000 samples with judges from other model families (Table 13). Three human raters also scored 150 samples on the seven dimensions: human–judge agreement on the keep/reject decision (κ=0.69) approaches inter-human agreement (κ=0.74), and the raters judged 93.3% of traces logically sound.

Table 13: Cross-family judge agreement with Qwen3-32B on 2,000 samples.

Judge	Spearman ρ	Keep/reject agreement	Cohen's κ	Retention rate
Qwen3-32B (reference)	–	–	–	81.7%
GPT-4o	0.83	91.4%	0.72	79.8%
Llama-3.3-70B	0.79	89.7%	–	–

B.3 AUDIO SYNTHESIS AND STATISTICS

User audio: MegaTTS3 with four reference voices from LibriSpeech-clean, 24 kHz. **Assistant audio:** Qwen3-Omni with one fixed voice ("Chelsie"), 24 kHz. **Hallucination filter:** audio with fewer than 1.5 or more than 5.5 words per second is rejected (skipped for utterances under five words). Durations range from 1.9 s to 148.5 s (5th–95th percentile 14.6–83.1 s).

Table 14: Per-source breakdown of THINKINGTALK.

Source	Count	Share	Reply (words)	Trace (words)	Duration (s)
SmolTalk2 system chat	23,247	88.1%	105.1	78.1	44.7
SmolTalk2 everyday conversation	1,673	6.3%	113.6	75.9	46.7
SoundMind	1,467	5.6%	44.2	69.8	20.4

B.4 EXAMPLE SAMPLES

Example A (SmolTalk2 system chat). *User:* "I'm trying to set up VoiceOver on my iPhone but I'm not sure where to start. Can you help me?" *Trace:* "The user needs help setting up VoiceOver on their iPhone. Main steps: navigate to Accessibility, enable VoiceOver, learn basic gestures. Structure step-by-step in a friendly tone…" *Reply:* "Sure thing! Let's get your iPhone's VoiceOver feature up and running. First, open the Settings app and scroll down to tap Accessibility. From there, select VoiceOver. Toggle the switch to On, and your iPhone will start reading the screen aloud…"

Example B (SoundMind). *User:* "Can you tell me what the Spanish word for hello is?" *Trace:* "The user is asking for the Spanish equivalent of hello. The most common term is hola; appropriate for most situations…" *Reply:* "The Spanish word for hello is hola. It's pronounced like oh-lah and works in most situations. If you want to be more formal…"

C Mechanism Probes: Protocols and Additional Results

C.1 Inference-Time Interventions

All interventions in Table 1 use the 8B model with a total budget of 512 denoising steps. *Think-first, frozen*: the reasoning segment is denoised to completion while reply positions stay masked, then the reply is denoised with the trace fixed. *Speech-first*: the reverse order. *Attention restrictions*: the named attention direction is masked at every layer and denoising step ("reasoning cannot attend to speech", "speech cannot attend to reasoning", or both), with the rest of the decoding procedure unchanged. *No reasoning segment*: the trace is omitted and the reply is denoised directly. The benefit of reasoning grows with question difficulty (+3.3 on LlamaQuestions, +5.7 on TriviaQA, +9.4 on WebQuestions), so the trace can be skipped on easy queries to save latency.

C.2 Trace Manipulations

To test whether answers consume the content of the trace, rather than a static answer prior distilled from the rewriting model, we freeze a manipulated trace and then denoise the reply conditioned on it (Table 15). Accuracy falls as the manipulation grows, and shuffled or irrelevant traces fall below having no trace, so the reply reads the trace's content. Because the rewriting prompt supplies the answer before the trace is written (Appendix B.1), traces are post-hoc; this experiment shows that the model uses them, not that they reflect how it computes the answer.

Table 15: Replies conditioned on frozen, manipulated traces (8B; S→S Avg, %).

Trace given to the reply	**S→S Avg**
Joint denoising (reference; not frozen)	50.6
Model's own trace, frozen (think-first)	47.8
Truncated to the first 25%	45.6
Shuffled	42.3
Irrelevant	41.0
No trace	44.4

C.3 Matched-Training Arms

The three arms of Table 2b are Stage-2 runs of the 1.7B diffusion variant with identical data, steps, and hyperparameters; only the corruption pattern differs. *Arm A (joint)* corrupts trace and reply together as in Equation (4) and decodes with Algorithm 1. *Arm B (think-first curriculum)* corrupts reasoning positions only in samples whose reply is fully masked and reply positions only in samples whose trace is clean; it is decoded by denoising the trace to completion before the reply, which is exactly its training distribution. *Arm C (trained-in restriction)* applies the text-only attention restriction for reasoning positions during training and at inference. All three arms are decoded with $T{=}512$ steps, and each arm was trained once. Arm A shares its nominal configuration with the "Diffusion, w/ ThinkingTalk" row of Table 2a, which was trained in an earlier, separate run; the two differ by 2.4 points on average (LlamaQ 42.5 vs. 42.3, TriviaQA 17.3 vs. 15.1, WebQ 27.6 vs. 22.7). We have no training-seed replicates with which to attribute this difference.

C.4 Error Injection

We select 300 examples (100 per S→S benchmark) that the 8B model answers correctly under default decoding. After 25% of the denoising steps, we force-commit one corrupted token in the reasoning segment that changes the key entity or number, and let decoding continue. We record (i) whether the final spoken answer is still correct; (ii) the KL divergence between the predictive distributions over still-masked reasoning positions with and without the injection, averaged over those positions; (iii) whether the argmax of any still-masked reasoning position changes within three steps; and (iv) for recovered runs, whether the later trace text explicitly overrides the corrupted value. The control repeats the injection while reasoning positions cannot attend to speech.

C.5 Decoding Instrumentation

On about 200 examples per benchmark we log, for every step of Algorithm 1, which positions are committed. *Commitment order*: the fraction of reasoning tokens committed after the step at which the first reply-speech token is committed. *Counterfactual flips*: after each speech commitment, we rerun the forward pass on the state with that token re-masked and compare the argmax and the distribution at every still-masked reasoning position; a *flip* is a change of argmax at one or more positions, and we average the KL divergence over those positions. Overall, 17.4% of speech commitments flip at least one reasoning argmax (12.1% with Fast-dLLM v2), 63% of examples contain at least one flip, and the mean KL is 0.031 with a heavy tail (top-decile mean 0.24). *Attention mass*: the share of reasoning-position attention mass that falls on committed reply-speech tokens is 2.9% in the first quarter of denoising steps, 18.4% in the last, and 11.7% averaged over all steps. *Commitment invariance*: across 2.1M logged commitments no committed token changed, as Algorithm 1 guarantees.

Missing controls. The flip statistic has no null condition. The natural null re-masks a just-committed *reasoning* token and measures flips at still-masked reasoning positions, and statistics restricted to the tokens that are finally committed would separate consequential flips from near-ties. We have not run either, so the counterfactual statistics describe how often speech commitments coincide with changes in reasoning predictions, not how much speech content drives them.

D Additional System-Level Results

D.1 Evaluation Protocol

S→S accuracy: the spoken reply is transcribed with Whisper-large-v3 and scored after standard text normalization, identical to Moshi, MinMo, and Qwen2.5-Omni. TTS WER: generated audio is transcribed with Whisper-large-v3 and compared with the input text. Latency: one H100, batch size 1, bf16, each system's official inference code, no serving-side optimization for any system. Cells marked † in our tables are copied from the cited source.

D.2 Full ASR and TTS Results

Table 16: ASR and TTS WER (%↓) on four benchmarks. '/': unsupported; '–': not evaluated; †: copied from the cited source. Qwen3-Omni (30B-A3B) is an above-scale reference.

	LS-Clean		**LS-Other**		**VoxPopuli**		**Common Voice**	
Model	ASR	TTS	ASR	TTS	ASR	TTS	ASR	TTS
Qwen2-Audio	1.8	/	3.5	/	7.1	/	8.6	/
Phi-4-Multimodal	2.1	/	3.6	/	6.3	/	9.2	/
Audio Flamingo 3	1.57†	/	3.13†	/	–	/	–	/
SpeechGPT	11.0	14.1	16.7	15.3	18.2	21.3	19.4	23.2
SpiritLM	6.0	6.7	11.0	9.5	14.3	19.4	15.4	22.4
Moshi	5.5	7.0	12.0	7.2	8.8	10.6	9.4	14.2
MinMo	1.8	6.7	3.9	7.5	6.7	10.9	8.0	13.5
Llama-Omni2	3.5	10.1	4.0	9.2	9.5	12.4	11.3	17.2
Qwen2.5-Omni	1.8	5.0	3.6	5.4	5.8	9.2	7.4	12.5
GLM-4-Voice	2.8	5.8	4.0	6.3	9.5	12.5	10.9	16.4
Step-Audio 2 mini	1.33†	4.3	2.86†	5.5	–	–	–	–
Qwen3-Omni (30B-A3B)	1.2	4.5	2.4	4.9	4.5	7.8	6.8	10.9
DiffuSpeech	3.0	6.2	4.3	6.8	7.1	10.3	7.6	11.7

D.3 Perceptual TTS Quality

We score LibriSpeech TTS outputs with UTMOS (Saeki et al., 2022) and measure speaker consistency as the cosine similarity of WavLM (Chen et al., 2022) speaker embeddings within each generated

response (Table 17). DIFFUSPEECH lies between the Moshi and MinMo tiers and below Qwen2.5-Omni. These are automatic proxies; we have no subjective MOS/CMOS or prosody evaluation.

Table 17: Automatic perceptual metrics on LibriSpeech TTS outputs.

System	UTMOS (clean)	UTMOS (other)	Speaker similarity
Qwen2.5-Omni	4.14	4.02	0.958
GLM-4-Voice	4.02	3.89	0.941
MinMo	3.97	3.85	0.946
DIFFUSPEECH	3.86	3.74	0.947
Moshi	3.71	3.58	0.928
Ground truth	4.09	3.92	–

D.4 TEXT BENCHMARKS WITH AR CONTEXT

Table 18 adds AR speech LMs to Table 5a. These rows reflect largely their different base LLMs; the relevant comparison for our claim is DIFFUSPEECH vs. LLaDA.

Table 18: Text benchmarks (T→T). AR rows are context only; '—': not evaluated.

Model	Type	MMLU	TriviaQA	GSM8K	MATH	HumanEval
Phi-4-Multimodal	AR	67.3	58.5	88.6	72.0	76.5
MinMo	AR	58.5	54.8	58.1	28.5	42.3
Moshi	AR	49.8	48.5	40.3	—	—
SpiritLM	AR	36.9	42.0	21.5	—	—
LLaDA (base)	Diff.	65.9	55.6	70.3	32.5	38.4
DIFFUSPEECH	Diff.	66.2	60.3	72.8	33.8	40.1

D.5 SPEECH REASONING (MMSU)

Table 19: MMSU (Wang et al., 2025a) via VoiceBench (Chen et al., 2024) (accuracy, %, S→T). Qwen3-Omni (30B-A3B) is an above-scale reference.

Model	Type	MMSU	Model	Type	MMSU
Qwen2-Audio	AR	35.7	Llama-Omni2	AR	33.4
Phi-4-Multimodal	AR	42.2	Qwen2.5-Omni	AR	61.3
SpeechGPT	AR	9.3	GLM-4-Voice	AR	30.5
SpiritLM	AR	14.5	Qwen3-Omni (30B-A3B)	AR	68.1
Moshi	AR	24.0	DiFFA	Diff.	29.6
MinMo	AR	43.2	DIFFUSPEECH	Diff.	39.0

On MMSU, DIFFUSPEECH (39.0) is well below Qwen2.5-Omni (61.3) and slightly below MinMo (43.2), but ahead of DiFFA (29.6). Its multiple-choice speech-understanding ability is therefore not a strength, consistent with Stage 2 targeting open-ended spoken QA.

D.6 VARIANCE AND PER-BENCHMARK DEGRADATION

Over five inference seeds, DIFFUSPEECH's S→S average is 50.6 ± 0.3. A bootstrap over test items gives a 95% CI of [8.3, 9.9] for its $\Delta_{T \to S}$ and [+3.6, +6.0] for its S→S advantage over Qwen2.5-Omni (+4.8). Training-seed variance is not available. Table 20 breaks $\Delta_{T \to S}$ down by benchmark.

Table 20: S→T-to-S→S degradation per benchmark (points; lower means more preserved).

Model	LlamaQ	TriviaQA	WebQ	Avg
Moshi	4.1	11.6	6.8	7.5
Llama-Omni2	9.6	11.4	3.3	8.1
Step-Audio 2 mini	7.6	11.6	9.8	9.7
Qwen2.5-Omni	8.3	12.8	11.1	10.7
MinMo	12.7	14.6	15.1	14.1
GLM-4-Voice	14.0	12.6	16.3	14.3
DIFFUSPEECH	3.5	11.9	11.8	9.1

D.7 INFERENCE LATENCY

Table 21: Wall-clock latency measured by us (one H100; batch size 1; bf16; official inference code; no serving optimization for any system; ∼5 s question, ∼10 s reply). AR rows are matched-harness context, not a ranking: optimized AR serving stacks would be faster.

System	Decoding	S→S Avg	TTFA (s)	Total (s)	RTF
DIFFUSPEECH	Vanilla diffusion (T=512)	50.6	3.2	18.5	1.85
DIFFUSPEECH	+ Fast-dLLM v2 (training-free)	49.7	1.0	6.2	0.62
Moshi	AR, streaming	45.1	0.2	8.5	0.85
MinMo	AR	43.1	0.4	9.3	0.93
Qwen2.5-Omni	AR	45.8	0.5	9.8	0.98

With Fast-dLLM v2 the per-benchmark S→S accuracies are 67.8 (LlamaQ), 32.9 (TriviaQA), and 48.4 (WebQ). Time to first audio grows with the reasoning prefix under block decoding, because reply speech can be emitted only after the blocks that contain the trace are decoded.

D.8 LENGTH SCALING

Table 22: Quality and latency vs. output length n (speech tokens) at fixed T=512 and at T=n. WER: long-form TTS on concatenated LibriSpeech passages. Quality: long-form instruction following scored 1–5 by an LLM judge. '=': identical to the T=512 column.

	WER ↓		Quality ↑		Total s (RTF)	
n (audio)	T=512	T=n	T=512	T=n	T=512	T=n
256 (10 s)	6.1	–	4.4	–	5.9 (0.58)	–
512 (20 s)	6.2	=	4.3	=	6.2 (0.30)	=
1024 (41 s)	7.3	6.9	4.1	4.2	6.8 (0.17)	12.9 (0.31)
2048 (82 s)	9.6	7.2	3.7	4.1	7.9 (0.10)	26.8 (0.33)

Quality decays once n exceeds T, because each step must commit $\lceil n/T \rceil > 1$ tokens. Setting T=n recovers most of the loss but makes total time grow linearly in n, so at matched long-form quality diffusion decoding scales like AR decoding up to a constant factor. The conversational regime of our benchmarks (median effective length 731 tokens) lies where the measured degradation is at most 1.1 WER points.

D.9 END-TO-END VS. CASCADE

A cascade transcribes the question with Whisper, answers with reasoning using LLaDA-8B, and synthesizes the answer with MegaTTS3. In the same harness it is less accurate than DIFFUSPEECH on every benchmark (Table 23). Plausible reasons are that ASR errors propagate into reasoning, that the text model sees only the transcript and none of the input audio, and that the three components are not optimized jointly; we did not isolate these factors.

Table 23: End-to-end DIFFUSPEECH vs. an ASR→LLM→TTS cascade (S→S accuracy, %; total latency on one H100).

System	LlamaQ	TriviaQA	WebQ	Avg	Total (s)
Whisper → LLaDA-8B → MegaTTS3 (cascade)	64.0	30.2	45.8	46.7	8.5
DIFFUSPEECH + Fast-dLLM v2	67.8	32.9	48.4	49.7	6.2
DIFFUSPEECH, vanilla T=512	68.5	33.5	49.7	50.6	18.5

D.10 DENOISING STEPS AND PRE-COMMITMENT DRIFT

Table 24 varies T with the target length fixed at n=1024, a stress test that differs from the task-specific $n_{\max}$ used for the main ASR/TTS results. ASR improves monotonically with T, TTS is best at T=512, and spoken QA is robust. Committed tokens never change under Algorithm 1. What changes with T is *pre-commitment drift*: a still-masked position's argmax can be correct at some step and change before the position is committed. The fraction of positions whose argmax was correct at some pre-commitment step but that were committed to a different value grows with T, since a larger T gives each position more steps at which to drift, while a small T commits more tokens per step and under-refines. We use T=512.

Table 24: Performance vs. denoising steps (n=1024). LlamaQ S→T accuracy (%↑); WER (%↓); drift: % of positions committed to a value other than an argmax they held earlier that was correct.

Steps T	LS-Clean TTS	LS-Clean ASR	LlamaQ S→T	Drift
1024	8.45	**3.08**	**72.13**	6.8
512 (default)	**6.20**	4.16	72.06	3.6
256	7.60	5.27	70.52	2.4

D.11 STAGE-2 TASK-PROPORTION SENSITIVITY

Varying the Stage-2 task proportions by 10–15 points changes LlamaQ S→S by at most 4.5 points and TTS WER by at most 0.9; a uniform mixture is within 1.7 points of the default (Table 25).

Table 25: Stage-2 task-proportion sensitivity (default in bold).

S2S	S2T	T2T	ASR	TTS	LlamaQ S→S	LS-Clean TTS
0.30	**0.30**	**0.20**	**0.10**	**0.10**	**68.5**	**6.2**
0.40	0.20	0.20	0.10	0.10	69.8	6.8
0.20	0.40	0.20	0.10	0.10	65.3	5.9
0.25	0.25	0.25	0.125	0.125	66.8	6.5

D.12 COMPARISON WITH DIFFA

Table 26: DiFFA vs. DIFFUSPEECH.

Aspect	DiFFA (Zhou et al., 2025b)	DIFFUSPEECH
Speech understanding (S→T)	✓	✓
Speech generation (TTS, S→S)	✗	✓
Reasoning traces	✗	✓
Speech tokens in vocabulary	input only	input and output
Corruption schedule	–	modality-specific (Table 10)
Training	single stage	two stages
Tasks	ASR, S→T QA	ASR, TTS, S→T, S→S, T→T
S→T QA Avg (Table 4)	45.9	59.6

D.13 PILOT: SPEECH INFILLING

Masked diffusion natively supports editing a span of generated speech in context. In a 50-sample pilot, we mask a 2-second span of a committed reply, edit the corresponding text, and re-denoise only that span conditioned on the surrounding speech. The infilled span has a WER of 6.9 (vs. 6.2 for full generation) and seam-region UTMOS of 3.83 (vs. 3.86). The pilot is small; style-controllable TTS and broader editing are left to future work.

E QUALITATIVE EXAMPLES

The traces below are verbatim model outputs. They illustrate the output format only: they show final states, not the order in which tokens were committed, and so do not by themselves show co-refinement.

Example 1: conversational QA (S→S).

User question (spoken)

"Can you give me a simple way to teach my new golden retriever puppy to sit when I say the word?"

Reasoning trace (text, internal)

The user wants a straightforward method to teach a golden retriever puppy to sit on command. Core steps: lure with a treat to guide the puppy into the position, introduce the verbal cue, reinforce through repetition. Puppies have short attention spans, so sessions must be brief. Avoid clickers / advanced techniques. Align with the oral, friendly style of the spoken reply.

Spoken answer (transcribed)

"Sure! Here's a simple way to teach your puppy to sit. First, get a treat your pup loves and stand in front of them. Hold the treat close to their nose and slowly move it up toward their head. As they follow the treat, their bottom will lower. The moment they sit, say 'sit' and give them the treat. Repeat this a few times, then take a break. Over time, say 'sit' before moving the treat, and eventually phase out the treat, using praise instead. Keep it short and fun—your pup will learn fast!"

Example 2: mathematical reasoning (T→T).

User question (text)

"Define $f(x) = 2x + 3$. Suppose that $g(x + 2) = f(f(x - 1) \cdot f(x + 1) + f(x))$. Find $g(6)$."

Reasoning trace (text, internal)

1. Goal: find $g(6)$*, given* $f(x) = 2x + 3$. *2. Set* $x + 2 = 6 \Rightarrow x = 4$. *3.* $g(6) = f(f(3) \cdot f(5) + f(4))$. *4.* $f(3) = 9$, $f(5) = 13$, $f(4) = 11$. *5. Argument:* $9 \cdot 13 + 11 = 128$. *6.* $f(128) = 259$.

Text answer

"Set $x+2 = 6$, so $x = 4$. Then $g(6) = f(f(3) \cdot f(5) + f(4)) = f(9 \cdot 13 + 11) = f(128) = 259$."