Conversational Speech Generation on an Encoder-Free Multimodal Backbone

Gianni Sanrochman
gsanrochman@proton.me

November 16, 2026

Abstract

Speech-capable language models usually bolt a separate audio encoder onto a text LLM. We ask a narrower question: is a frozen multimodal LLM's *own*, encoder-free audio pathway enough, and can lightweight adapters alone give it a voice? We couple a frozen Gemma-4 backbone with a CSM-style speech head, a semantic-codebook predictor plus a depth transformer over Mimi residual-vector-quantized codes, and train only QLoRA adapters and the head on a single 16 GB GPU. **(RQ1)** On the smaller E2B backbone the recipe teaches a fully frozen, quantized LLM to speak: **13% word error rate, 88% word recall** on held-out text-to-speech. **(RQ2)** For the encoder-free 12B backbone we characterize *how* audio must reach it. (i) Naive 4-bit NF4 quantization *silently destroys the backbone's audio comprehension* (held-out transcription word recall falls from **0.65 to 0.12**) while sparing prosody; keeping only the audio tower and projector in bf16 (a ∼10.5 GB hybrid load) restores it. (ii) *Delivery matters*: the native multimodal pathway conveys lexical content, whereas a hand-built feature-splice of the same features conveys prosody but not content. Exploiting these, a two-stage cascade (native understanding into the frozen backbone, then the speech head) gives *conversational speech-to-speech with no backbone training*, rendering held-out replies at **13.1% WER / 91.9% recall**. (iii) But prosody-conditioning, replying based on *how* the query was said, is not yet realized: output prosody shows only a weak loudness echo, with no pitch or timing tracking. Two methodological findings recur: teacher-forced codebook cross-entropy is *not* a proxy for free-running intelligibility (a memorizing model scores well yet decodes at 141% WER), so word error rate on decoded audio is the necessary gate; and the Mimi codec imposes an entropy floor that cross-entropy cannot cross. We further collapse the cascade into a *single* end-to-end model that delivers audio natively and conditions the speech head on the backbone's latent reply: it speaks intelligible, grounded, full-length replies in one forward pass (WER 0.00 on the best held-out clips). A controlled ablation, however, shows heavy adapter training on limited data erodes the backbone's out-of-distribution competence: the frozen backbone's intelligence stays intact (disabling the adapter restores it), but the adapter masks it, making robustness a data-and-adaptation problem rather than an architectural one. Code, checkpoints, and data are released.

1 Introduction

Multimodal language models are usually extended to speech by attaching a separate audio encoder, typically a Whisper variant, to the language model and projecting its features into the token stream. That recipe is effective but heavy: it adds an encoder, treats audio as a foreign modality to be translated in, and is rarely trainable on a single consumer GPU. Recent backbones such as Gemma-4 are instead *encoder-free*: they ingest audio through the model's own multimodal pathway with no external encoder, which sharpens the question. If the backbone already *understands* audio

natively, is that pathway enough to also give it a *voice* using only lightweight adapters: no encoder, no backbone fine-tuning, on 16 GB? We keep the language model frozen, keep its native audio pathway, and learn only a small speech-output head plus low-rank adapters.

Research questions.

- **RQ1 (foundation).** Can lightweight adapters teach a frozen 4-bit LLM to generate intelligible speech at all? We answer yes on the E2B backbone (§5.1).
- **RQ2 (thesis).** Using the backbone's *encoder-free* audio input, can the model produce a spoken response conditioned on *how* the query was said (prosody, emotion), not merely what was said?

A negative RQ1 would not damage RQ2; a positive RQ1 is independently useful, a cheap way to give an already-deployed frozen LLM a voice.

Contributions.

1. An encoder-free recipe coupling Gemma-4's native audio features with a CSM-style speech head (semantic codebook plus a depth transformer over Mimi RVQ codes), trainable end-to-end under QLoRA on 16 GB.
2. Evidence for RQ1: a fully frozen, quantized LLM made to speak intelligibly (13% WER / 88% recall) by adapters alone (§5.1).
3. A quantization finding: 4-bit NF4 quantization of the *whole* backbone silently destroys its audio comprehension while sparing prosody; protecting only the audio tower and projector in bf16 restores verbatim understanding at ∼10.5 GB (§5.2.1).
4. A delivery finding: the backbone's native multimodal audio pathway conveys lexical content, whereas a hand-built feature-splice of the same features conveys prosody but not content, isolating audio *delivery*, not quantization or capacity, as the bottleneck (§5.2.2).
5. A working demonstration of RQ2: a two-stage cascade (native understanding → frozen-backbone CSM speech head) holds a relevant spoken conversation with *no backbone training*, rendering held-out replies at **13.1% WER / 91.9% recall**; a head-refinement run fixes word-onset dropping (§5.2.4). Prosody conditioning, replying to *how* the query was said, is *not* yet realized: output prosody shows only a weak loudness echo, a clean negative bound (§5.4).
6. A *single* end-to-end model collapsing the cascade: native audio delivery with the speech head conditioned on the backbone's own latent reply, speaking intelligible, grounded, full-length replies in one forward (WER 0.00 on the best held-out clips). A three-way ablation localizes its one weakness, out-of-distribution template collapse, to adapter overfitting, not the architecture: the frozen backbone's intelligence stays intact and is restored by disabling the adapter. A hybrid that ablates only the content-selection pass then voices the reply with the adapter on recovers correct content with largely intelligible speech, with no retraining (§5.3).
7. A methodological finding: teacher-forced codebook cross-entropy (c_0) is not a proxy for free-running intelligibility (and the same trap recurs for comprehension); word error rate on decoded audio is the necessary gate (§5.6).

8. A characterization of the Mimi codec entropy floor that caps naturalness under cross-entropy training (§5.7).

2 Related Work

Speech and audio LLMs. Most speech-capable LLMs attach a separate audio encoder to a text LM. On the understanding side, SALMONN [16] fuses Whisper [13] and BEATs encoders into a frozen LLM via a Q-Former, and Qwen-Audio [3] couples a Whisper-style encoder to an LLM; both are speech-in, text-out. For speech *generation*, a dominant line discretizes audio into units or codec tokens: GSLM [10] models raw-audio units, VALL-E [17] is a codec-token LM for text-to-speech, and SpeechGPT [21] and AudioPaLM [14] extend an LLM's vocabulary with speech units and fully (re)train. Full-duplex models such as Moshi [4] train a bespoke speech backbone with the Mimi codec, while LLaMA-Omni [6] adds a speech adaptor and streaming decoder atop an LLM but still bolts a Whisper encoder on the input. We differ on four axes: (i) no external audio encoder, we use Gemma-4's *native* audio pathway; (ii) the backbone stays frozen, adapted only by QLoRA plus a speech head on a single 16 GB GPU; (iii) bidirectional conversational speech; and (iv) we gate on free-running WER rather than teacher-forced loss (§5.6).

Frozen-backbone and text-free speech-to-speech. Three recent systems bracket our setting. Freeze-Omni [18] also keeps the LLM frozen, but, like our cascade, routes a streaming speech *encoder* into the LM on input and drives an autoregressive speech decoder from the LM's *textual* response on output: a frozen-LLM, text-mediated pipeline whose stated motivation, avoiding catastrophic forgetting under limited data, is exactly ours. Qwen2.5-Omni [19] removes that text bottleneck with its Thinker–Talker design, the Talker "directly utilizes the hidden representations from the Thinker to produce audio tokens", but trains *end-to-end* rather than freezing the backbone. Concurrent with our work, MOSS-Speech [22] reaches true text-free speech-to-speech on an initially frozen backbone by carving modality-specific layers into the transformer itself: of Qwen3-8B's 36 layers the first 32 form a shared trunk and the last four split into parallel text- and speech-prediction stacks, trained with the backbone frozen before a later partial unfreeze. Relative to all three, our study occupies a distinct point: native *continuous*-audio input (no discrete speech tokenizer and no external encoder), a backbone that stays *fully* frozen, and training confined to QLoRA plus a small external head on a single 16 GB GPU. We obtain the conversational loop through the same frozen-LLM, text-mediated cascade as Freeze-Omni, and report an honest negative result on the naive single-model bolt-on (§5.2.4); the hidden-state conditioning of Qwen2.5-Omni and the in-trunk modality split of MOSS-Speech both indicate why it failed and chart the route for collapsing our cascade into one network (§6).

Parameter-efficient adaptation of frozen models. We freeze the backbone and train only adapters: LoRA [9] (here in its 4-bit QLoRA form [5]) and bottleneck adapters [8]. This follows the "freeze the LLM, add a small bridge" paradigm of Flamingo [1] and LLaVA [11], but for audio in *and* out and *without* a separate modality encoder.

Conditioning frozen LMs on new modalities. Closest to our setting, Penov [12] adapt a frozen small language model to sensor streams (vision, audio, IMU) via conditional LoRA bridges: a hypernetwork generates LoRA weights at runtime from a frozen sensor encoder's features. We share the premise of adapting a frozen quantized LLM to a non-text input without retraining the backbone, but differ in mechanism and goal: (i) we use the backbone's native, encoder-free audio

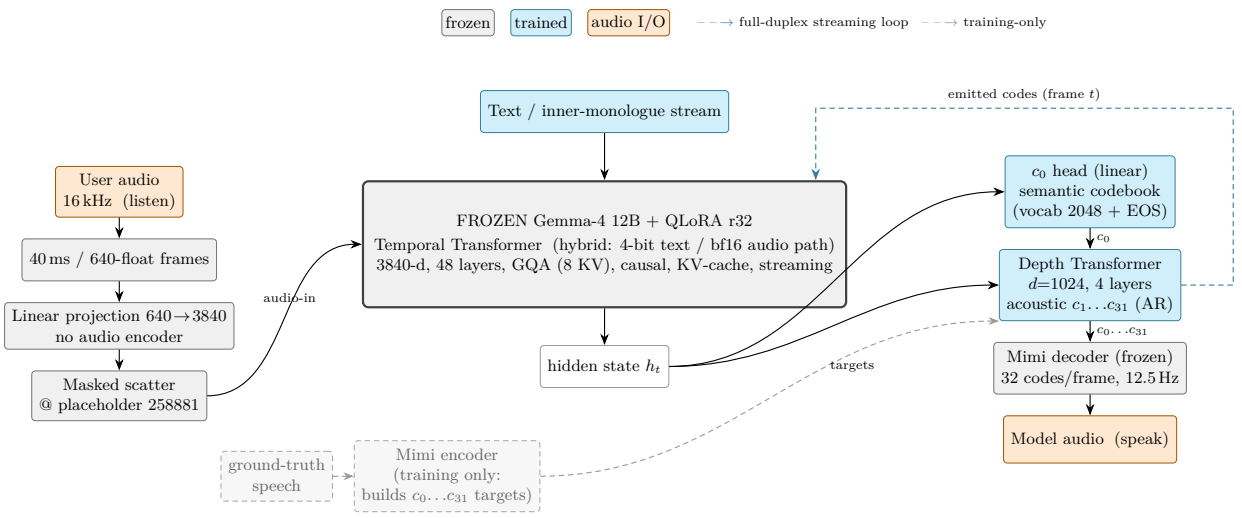


Figure 1: Architecture of the full-duplex conversational model. The frozen Gemma-4 12B decoder (the *Temporal Transformer*) is the only large component; the trainable parameters are the QLoRA adapters, the audio-frame embedding tables, the c_0 head, and the depth transformer. The understanding (listen) path is *encoder-free*: incoming audio is sliced into 40 ms / 640-float frames, linearly projected $640 \to 3840$, and scattered into the decoder at audio placeholder tokens, with no ASR/audio encoder. The generation (speak) path emits Mimi codes, a linear c_0 head for the semantic codebook and a depth transformer for the 31 acoustic codebooks, which the frozen Mimi decoder renders to waveform; *at inference no encoder runs on audio.* The Mimi encoder is used only at training time, to build the $c_0 \ldots c_{31}$ targets. The dashed loop is the full-duplex streaming path: the model's emitted codes at frame t are embedded as audio-out input at frame $t{+}1$, so listening and speaking proceed frame-synchronously. This is the *target* single-model layout; we demonstrate both the two-stage cascade of §5.2.4 (native-pathway understanding → this speak head) and the unified end-to-end model of §5.3, audio delivered through the native pathway inside one network, the head conditioned on the backbone's latent reply. The understanding path requires the audio tower and projector in bf16 (§5.2.1); only the language layers are 4-bit. Full duplex is the streaming extension of the `[text | audio-in | audio-out]` layout.

pathway rather than a separate sensor encoder feeding a weight-generating hypernetwork; (ii) our adapters are fixed LoRA plus dedicated speech-output heads, not input-conditioned generated weights; and (iii) we target bidirectional conversational speech (audio in and out). Methodologically their study and ours converge on the same warning (§5.6): an aggregate language-modeling loss does not measure the capability of interest, and a task-grounded probe is required.

3 Method

3.1 Encoder-free backbone

Figure 1 shows the full architecture. We build on Gemma-4 12B "Unified" [7], Google's encoder-free multimodal variant: 16 kHz audio is sliced into 40 ms / 640-float frames and linearly projected into the decoder's hidden space ($640 \to 3840$, one-to-one in time), merged at audio placeholder tokens (token id 258881) via a masked scatter into the input embeddings. There is no separate ASR/audio encoder; we access these projected features through the model's `get_audio_features`.

The smaller E2B/E4B audio variants instead use a USM conformer encoder and are *not* encoder-free; our encoder-free claim is specific to the 12B. The 12B decoder is 3840-dimensional, 48 layers, grouped-query attention (8 KV heads), with no per-layer-embedding tables and no KV-sharing.

3.2 Speech-output head

Audio is represented as Mimi codes [4]: a streaming split-RVQ codec at 12.5 Hz whose released checkpoint provides 32 codebooks (1 semantic, 31 acoustic). Following the CSM head topology [15], a linear semantic head predicts codebook 0 (c_0, vocabulary 2048 plus an end-of-speech token) and a depth transformer (a causal transformer, d=1024, 4 layers, 8 heads) autoregressively predicts the remaining 31 acoustic codebooks conditioned on the pre-frame hidden state and the sampled c_0. The loss is $\mathcal{L}_{c_0} + \lambda \mathcal{L}_{\text{depth}}$ with the depth term computed on a random subset of frames per step (we use 1/16 to 1/4). Each audio frame's input embedding is the sum of its 32 codebook embeddings (new tables, untied, unlike CSM). Unlike CSM (a from-scratch Llama backbone) we attach this head to the frozen Gemma-4 decoder. The sequence layout is `[text | audio-in | audio-out frames]`; the hidden state at the last text position predicts the first output frame, and frame t predicts frame t+1 (shifted next-frame prediction).

3.3 QLoRA training on 16 GB

The backbone is frozen in 4-bit NF4 with double quantization and bf16 compute; the Mimi codec is frozen. The trainable parameters are the LoRA adapters [5] (rank 32, applied to all linear layers, attention and MLP, following a coverage-over-rank principle), the audio-frame embedding tables, the c_0 head, and the depth transformer. We optimize with PagedAdamW8bit, batch size 1 with gradient accumulation 16, gradient checkpointing, and a per-clip frame cap. We sample at evaluation with temperature 0.6–0.9 and top-k 50 (cooler settings yield longer intelligible output).

3.4 Curriculum: Phase A then Phase B

Phase A (text→speech) teaches the head to speak on clean read speech; Phase B turns the audio input on and learns understanding-conditioned response on conversational query/response audio. The order is forced: Phase B resumes from a model that can already speak, so audio-conditioning has speech to attach to. A key lesson (§5.6) is that Phase A needs the *volume* of a large clean corpus (LibriTTS), not a small conversational set; the latter merely memorizes.

4 Experimental Setup

Hardware. All training runs on a single 16 GB consumer GPU (∼15.4 GB usable). The frozen 12B loads at ∼10.6 GB; capped-sequence audio-in training fits at ∼13.7 GB. The decisive memory levers are batch-1 with gradient accumulation, SDPA attention (eager materializes $O(\text{seq}^2)$ fp32 attention), a headless display, and `max_split_size_mb:256`; note that `expandable_segments` is incompatible with bitsandbytes 4-bit kernels and silently no-ops. A single ∼$7 rented-A100 session was used for one larger E2B run; that workload is memory-bound, so the A100 gave only ∼1.5× the consumer-card throughput, not the nominal compute ratio.

Data. Phase A trains on a ∼110 h / 90.6k-clip subset of LibriTTS [20] (24 kHz, CC BY 4.0). Phase B uses the English split of OpenS2S [2] (Apache-2.0): 49,368 paired (audio-query, audio-response) examples with transcripts and emotion labels. We re-encode all response audio with

Mimi at 12.5 Hz (OpenS2S itself ships raw WAVs tokenized with GLM-4-Voice, which we do not use).

Evaluation. We decode free-running (the model consumes its own sampled frames), transcribe with faster-whisper, normalize both reference and hypothesis with the standard English text normalizer, and report WER and *word recall* (the fraction of reference words present in the hypothesis, a multiset measure that is immune to the insertions that drive WER above 100% at low intelligibility). Double-blind human listening is the gold reference; it agrees closely with the recall harness.

5 Results

5.1 RQ1: adapters teach a frozen LLM to speak

We establish RQ1 on the smaller E2B backbone (text→speech; with no audio input involved, the encoder-free distinction does not apply here).

Intelligibility (M3). Trained on ∼840 h (LibriTTS-R plus LJSpeech), the frozen 4-bit LLM with QLoRA adapters and the speech head reached a final-checkpoint mean battery word-recall of **71%**, with multiple novel sentences at 100% recall (e.g. "Please close the door before you leave"); train/val c_0 were 2.54/2.65 with no overfitting gap. To our knowledge this is the first demonstration of intelligible speech from a fully frozen, quantized LLM via parameter-efficient adaptation alone, produced on a single 16 GB GPU plus a ∼$7 rented-A100 session.

Duration (M4). A same-speaker concatenation phase (mean ∼18 s clips) cured early truncation: speech sustains ∼11–13 s at 86–90% voiced. The final 10-clip battery scored **mean WER 13%, word-recall 88%** (novel sentences 91%/9%; reused probes 84%/17%), with $c_0 \approx 2.15$. Residual errors are acoustic, not lexical ("lazy goat"/dog).

Human–machine agreement. Double-blind human transcription tracked the recall harness closely (human 14–22% vs. machine ∼18% in the M2 era), supporting free-running WER/recall as the quality gate rather than teacher-forced loss.

5.2 RQ2: from naive integration to a working cascade

The 12B encoder-free pipeline is validated end-to-end: a training step consumes `[text | audio-in | audio-out]` with audio features from `get_audio_features` and propagates gradients at a 13.7 GB peak, and a text-conditioned speech head trained this way voices supplied replies near verbatim. But making the frozen backbone respond to *what the query said* (content), not only *how* (prosody), exposed two failure modes that we had to isolate before a conversational system worked.

5.2.1 4-bit quantization destroys audio comprehension

The RQ1 recipe loads the entire backbone in 4-bit NF4. Probing the 12B's audio understanding through its *native* multimodal chat-template, the 4-bit model transcribes nothing and answers generically (e.g. "You haven't said anything yet") regardless of the spoken input; the same probe in full bf16 transcribes verbatim ("The quick brown fox…"; a spoken name and a spoken question correctly answered). The cause is that bitsandbytes quantizes the audio tower and audio projector along with the language layers, while the library's automatic skip list protects only tied/output embeddings. Listing the audio tower, `embed_audio`, the vision modules, and `lm_head`/`embed_tokens` in `llm_int8_skip_modules` keeps the audio pathway in bf16 (a ∼10.5 GB "hybrid" load that still

Table 1: Does the frozen backbone receive the spoken *content*? Free-running transcription of 40 held-out OpenS2S queries (normalized), through the native multimodal pathway, at bf16 vs. naive whole-model 4-bit NF4. Word recall is the fraction of reference words present in the output.

delivery	precision	word recall	WER
native pathway	bf16	**0.65**	0.62
native pathway	4-bit NF4 (naive)	**0.12**	1.34

fits 16 GB) and restores verbatim understanding. Notably, *prosody* conditioning survived 4-bit (RQ2's original framing worked), but lexical content did not, so every earlier "comprehension" failure was this single artifact. Table 1 quantifies it: on 40 held-out OpenS2S queries, asking the model to transcribe the audio recovers 65% of content words in bf16 but only 12% in naive 4-bit, where the model typically refuses ("I cannot provide a transcription because there is nothing…"). We report word recall (content overlap) as the robust metric because the bf16 model paraphrases rather than transcribes verbatim, inflating its WER; the recall gap isolates the quantization effect.

5.2.2 Delivery mechanism: native pathway vs. feature-splice

Even with the audio path in bf16, *how* features enter the decoder decides whether content survives. We compare two deliveries of the identical audio features: (a) our hand-built path, which calls `get_audio_features` and splices the projected frames into `[BOS | audio | text]` (used since Phase B, and which carries prosody); and (b) the processor's native chat-template, which inserts audio soft-tokens and merges them by masked scatter. The splice path produces hallucinated transcripts, the model invents a fluent, unrelated query and answers *that* (free-running transcript WER $\sim$1.1), whereas the native path transcribes verbatim. Having eliminated training instability (§5.5), quantization (§5.2.1), and capability (the native path understands), we isolate audio *delivery* as the bottleneck: the manual splice transmits prosody but not lexical content.

5.2.3 End-to-end "inner monologue" attempts fail informatively

We first tried a single model that both understands and speaks. A *joint* layout `[audio | reply-text | reply-audio]` with loss on text and audio produced on-style but generic, off-topic replies: because the reply text is always present in the prefix (no text dropout), the objective rewards *vocalizing a supplied answer*, never *deciding* what to say, visible only under a text-free decode, not in the loss. A *grounded* variant adding a comprehension loss (transcribe the query, whose transcript already ships in the data) drove teacher-forced comprehension loss to $\sim$0.5 while free-running transcript WER stayed at 1.07–1.13: the model hallucinates a fluent unrelated query and replies to it, the c_0-is-not-intelligibility trap (§5.6), now recurring for *comprehension*. These negatives motivate decoupling understanding (native pathway) from speaking (the validated head).

5.2.4 A working cascade, and head refinement

Combining the two diagnoses gives a system that needs *no backbone training* (Figure 2). Both stages are the *same* frozen backbone, one set of weights, invoked twice, so the cascade is already a single model in *parameters*; "two-stage" names the information path, two forward passes joined by a discrete-text handoff. The "single end-to-end model" we target therefore removes not a second network but that text bottleneck: one forward pass in which the speech head conditions on the hidden states the backbone produces while natively ingesting the query, so prosody and

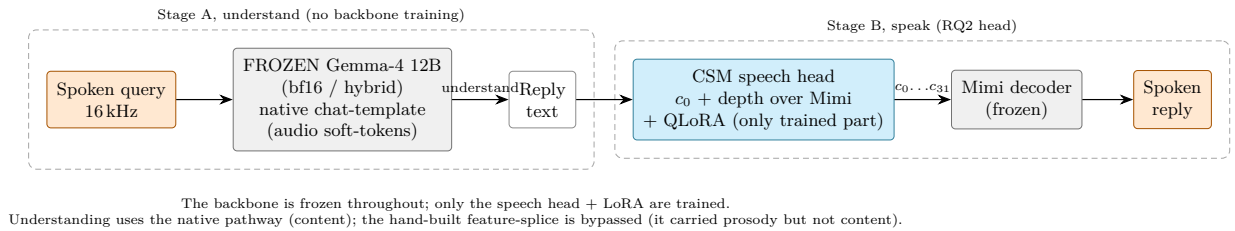
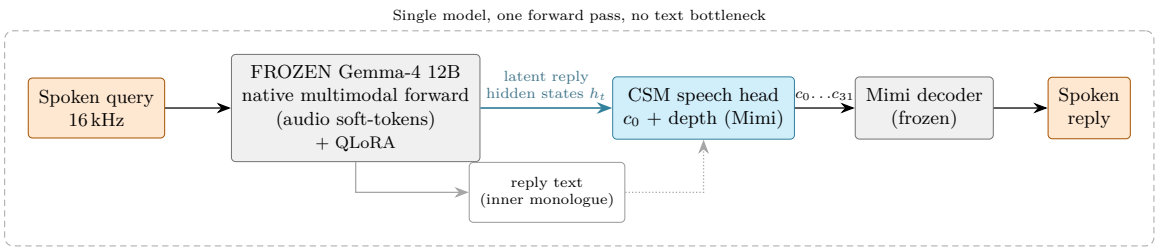


Figure 2: The two-stage conversational cascade (§5.2.4). Stage A feeds the spoken query to the frozen backbone through its *native* multimodal pathway and reads out a reply in text; Stage B is the CSM-style speech head (the RQ1 text→speech head) that voices the reply, rendered to waveform by the frozen Mimi decoder. The backbone is frozen throughout and the hand-built feature-splice (§5.2.2) is bypassed for understanding; only the speech head and LoRA are trained. Collapsing this into one network via native delivery is the remaining step.

Table 2: Cascade conversational battery: free-running output transcribed with faster-whisper and compared to the Stage-A reply text. Understanding (Stage A, full-bf16 native pathway) is held fixed; only the Stage-B head varies. *Onset* = leading word(s) rendered; *verbatim* = full reply with no word error. This per-clip battery illustrates the failure modes; the aggregate held-out result (40 queries: 13.1% WER, 91.9% recall) is reported in the text.

Battery	Stage-B head	onset	verbatim	residual
short (3, ∼3 s)	pre-refinement	1/3	0/3	dropped 1–3 leading words; one stutter
short (3, ∼3 s)	refined	3/3	3/3	,
long (3, 13–20 s)	refined	2/3	2/3	one sentence-initial substitution
OOD (unseen self-intro)	refined	1/1	1/1	rare-word subs. when long

paralinguistics can survive into the reply rather than being discarded at the text interface. Stage A is the frozen backbone in bf16/hybrid, fed audio through the native chat-template, emitting a reply *text*. Stage B is the CSM-style speech head (the RQ1 text→speech head) voicing that reply. On a small conversational battery the cascade closes the loop on-topic and intelligibly (a promotion query → a congratulation; "I've been feeling overwhelmed" → an empathetic reply), bypassing the broken splice for understanding while reusing the proven head for voice. A short head-refinement run (8k steps, warm-started, on the hybrid backbone) then removed a residual artifact in which the head dropped the first one to three words: clean verbatim rendering rose from 1/3 to 3/3 of the battery, and output extended to clean ∼13–20 s multi-sentence replies; an out-of-distribution probe (a spoken self-introduction absent from training) was understood and voiced verbatim, with only an occasional sentence-initial word substitution remaining on rare words. Across this run teacher-forced validation loss was flat while free-running quality jumped from 1/3 to 3/3, the c_0 trap once more. Table 2 summarizes the battery. At scale, on 40 held-out OpenS2S queries, the cascade renders the Stage-A reply at **13.1% WER and 91.9% word recall** (free-running, faster-whisper-normalized), matching the RQ1 text-to-speech intelligibility (§5.1).



The frozen backbone reasons in latent space; the head voices its hidden states directly (Thinker→Talker), so prosody can survive into the reply.
Only the head + QLoRA train, but the QLoRA that wires the head also perturbs the backbone's reasoning out-of-distribution (disabling it restores correct replies).

Figure 3: The unified single model (§5.3). The frozen encoder-free backbone ingests the spoken query through its *native* multimodal forward and the CSM speech head is conditioned on the *hidden states* of the autoregressively generated reply, the model's own latent reasoning, rather than on re-encoded text, so understanding and speaking happen in one forward pass with no text bottleneck (Qwen-style Thinker→Talker). Only the head and QLoRA are trained; the backbone is frozen. The honest caveat (§5.3): the QLoRA that wires the head onto the latent also perturbs the backbone's reasoning out-of-distribution, which a three-way ablation localizes (disabling the adapter restores correct replies).

5.3 Collapsing the cascade: a demonstrated unified model and its limit

We then built the single end-to-end model the cascade points to. `forward_unified` delivers the query audio through the backbone's *native* multimodal forward (`<audio>` placeholders + `input_features`, the path proven to carry lexical content in §5.2.2) and conditions the CSM head on the hidden states of the autoregressively generated reply, one forward pass, no text re-encoding, in the spirit of Qwen2.5-Omni's Thinker–Talker [19]. Only LoRA and the head are trained (backbone frozen); the head is warm-started from the cascade's and the model is trained on the full English OpenS2S split (4,898 pairs, 3,000 steps, on a rented A6000 because the gradient-checkpointing-off native forward exceeds 16 GB at full reply length).

It works: single-pass, intelligible, grounded. On held-out OpenS2S queries the unified model produces *full-length* (9–12 s) spoken replies that free-running faster-whisper recovers near-verbatim (**WER 0.00** on the best clips, 0.03–0.29 typical) in a single forward, with no text bottleneck. A zeroed-audio ablation, replacing the input features with zeros, changes the reply every time, confirming the output is grounded in the audio and not a learned prior. This is the collapse of the two-pass cascade into one network, achieved: the frozen encoder-free backbone hears the query, forms the reply in latent space, and the head voices that latent directly.

But adapter training erodes the backbone's general competence. The same model fails on out-of-distribution prompts. Asked "how are you?", "who are you?", or a factual question about a Google model, it returns OpenS2S *templates* (empathy boilerplate, "a picnic in the park", "Congratulations!") instead of engaging the backbone's knowledge. A three-way control (Table 3) isolates the cause. *Native Gemma*, the bare frozen model with no LoRA and no head, the same audio through its own multimodal chat-template, answers all three correctly, recovering the spoken name and giving accurate identity and factual replies, so the frozen backbone's intelligence is fully *active* and comprehends the speech with zero training. Disabling the LoRA on the trained model (*unified*−LoRA) likewise *restores* correct, knowledgeable replies. Only the trained *unified*+LoRA

Table 3: Backbone-activation control on spoken out-of-distribution queries. *Native Gemma*: bare frozen multimodal model (no LoRA, no head), audio through its own chat-template. *Unified*+LoRA: the trained single model. *Unified*−LoRA: the same model with adapters disabled. Native Gemma and unified−LoRA both answer correctly (backbone intelligence intact); only the trained adapter collapses to training-set templates. Replies abbreviated.

Spoken query	Native Gemma (no LoRA/head)	Unified +LoRA (trained)	Unified −LoRA (ablated)
"How are you?"	"I'm an AI… here is how I work…"	"You're not alone in feeling this way…"	"I'm doing well, thank you! How are you?"
"My name is Gianni. Who are you?"	"Hello, Gianni. I am a large language model trained by Google…"	"how about a picnic at the park…"	"I am an AI assistant… trained by Google… How can I help, Gianni?"
"Tell me about Google's model."	"… a family of multimodal LLMs… trained natively, not bolted on…"	"Congratulations! That's exciting news!"	"… Google's multimodal model… text, images, video, audio, code…"

collapses to templates. The backbone's latent reasoning was therefore intact throughout; the heavy all-linear LoRA, fit to ∼5k narrow examples, overrode it on inputs outside that distribution.

Speech-to-speech in the ablated state, and a partial fix for free. Table 3 reports the *replies*; what does the ablated model *sound* like? We voiced all three configurations end-to-end (audio in, audio out), adding a fourth query, "congratulate me, I got promoted" (Table 4). The two columns are mirror images. *Unified*+LoRA voices its reply near-perfectly (WER 0.00–0.12), but the reply is the wrong out-of-distribution template: told of good news, it offers to "talk about it". *Unified*−LoRA says the right thing yet cannot voice it (WER 0.94–1.00); the head was trained to ride full-strength-LoRA hidden states, and the ablated states are off its manifold. Content and intelligibility want opposite adapter strengths.

A *hybrid* inference path recovers much of the gap with no training, by splitting the single forward at its two sub-passes. We *think* (deliver the audio, decide the reply text) with the adapter *off*, so the pristine backbone chooses correct words, then *voice* (the head generates frames) with the adapter *on*, teacher-forcing that reply so the head sees in-distribution hidden states (Fig. 4). This halves to quarters the WER, and for a reply close to the training distribution ("Congratulations on your promotion") yields correct content *and* intelligible speech at once (WER 0.22). The residual error is itself diagnostic: it orders by how far the *correct* reply sits from the head's narrow training content (congratulation 0.22 < "I'm doing well" 0.30 < a technical model description 0.42). Two overfits stack, the LoRA narrowing *what the model says* and the head narrowing *what content it can voice*, both to the ∼5k OpenS2S examples; the hybrid removes the first at inference time, only broader-data retraining removes the second.

The trade-off. This sharpens cascade-vs-unified into a genuine tension. The cascade keeps the backbone's reasoning *untouched*, only the head is trained, preserving broad competence (91.9% recall) at the cost of a text bottleneck that discards prosody. The unified model removes that

Table 4: Speech-to-speech audio quality in the ablated state. Free-running faster-whisper WER of the generated audio against that configuration's own reply text (lower is better; ∼1.0 is unintelligible). *+LoRA*: the trained unified model, clean audio but out-of-distribution *template* content. *−LoRA*: adapters disabled, correct content but unvoiceable. *Hybrid*: think with adapters off, voice the chosen reply with adapters on. Native Gemma has no head, hence no audio.

Spoken query	+LoRA (wrong content)	−LoRA (ablated) (right content)	Hybrid (right content)
"How are you?"	0.05	1.00	0.30
"My name is Gianni. Who are you?"	0.12	0.94	0.59
"Tell me about Google's model."	0.05	1.00	0.42
"Congratulate me, I got promoted."	0.00	0.98	0.22

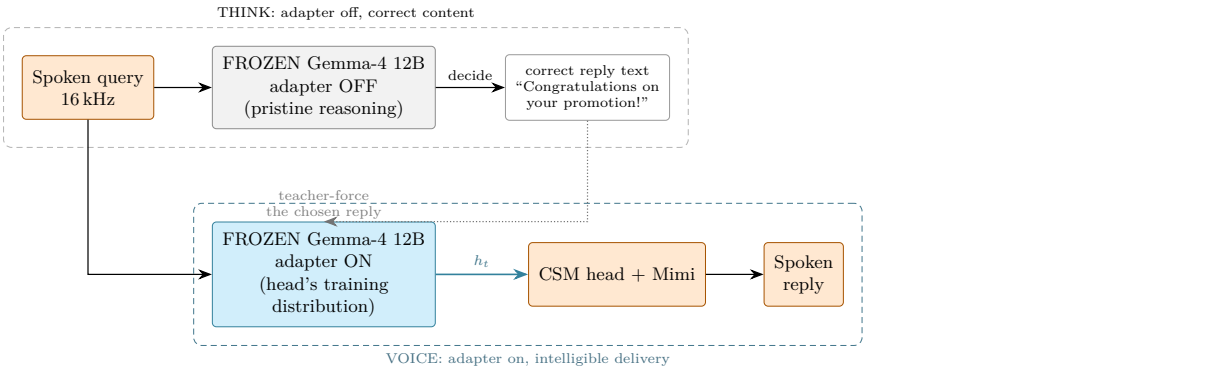


Figure 4: The hybrid inference path. Ablate only the *content-selection* sub-pass (adapter off, so the frozen backbone reasons correctly), then voice the chosen reply through the *delivery* sub-pass with the adapter on, on the head's training manifold. No weights change; it recovers correct content with largely intelligible speech where the reply resembles the head's training content.

bottleneck (one latent-conditioned forward, prosody-capable) but the adapter needed to wire the head onto the latent perturbs the very reasoning we want to voice. The remedy is to touch the "thinker" as little as possible: more and broader data, lighter or no LoRA, or the in-trunk modality layers and limited unfreeze of MOSS-Speech [22]. The core claim, *the model's own latent reasoning, spoken aloud, from a single frozen encoder-free backbone*, is demonstrated; making it robust out-of-distribution is a data-and-adaptation problem, not an architectural one.

Realizing the *single* end-to-end model hinged on getting delivery right. A first attempt, wrapping the hand-spliced audio embeddings in the chat-template framing tokens and training the joint objective end-to-end, does *not* suffice: although the teacher-forced text loss falls (to 2.3 held-out), the free-running model *bypasses the audio entirely.* Greedy decoding emits an identical reply across all queries, and an ablation that zeros the audio leaves that reply unchanged, so the response is independent of the input (the teacher-forced trap of §5.6 recurring at the system level). This confirms that true native delivery requires the backbone's own multimodal forward (audio soft-tokens placed by its `masked_scatter`), not framing wrapped around a manual splice. The working unified model of §5.3 builds on exactly this: it delivers audio through the native forward and conditions a warm-started head on the resulting hidden states, which, unlike the splice, suffices to produce intelligible, grounded, single-pass speech.

5.4 Prosody conditioning: a weak loudness echo only

RQ2 in its original form asks whether the spoken reply tracks *how* the query was said. We test this directly with a minimal-pair probe on the splice path (the path that carries prosody): for 12 held-out queries we hold the reply text fixed, manipulate *only* the input prosody into an "aroused" variant (higher pitch, faster, louder) and a "subdued" one (the inverse) via DSP, generate $K{=}3$ samples per variant, and measure whether the output's F0, pitch range, intensity, and duration shift in the input's direction, scored per-clip against the temperature-0.7 sampling noise. The result is largely *null*: output *intensity* weakly but consistently tracks input intensity (10/12 clips in the expected direction, $p \approx 0.02$; mean $\Delta{+}0.96$ dB), while *F0, pitch range, and duration do not track at all* (mean deltas below the sampling-noise floor, directional counts at chance). So although the splice delivers prosody into the backbone, the text-conditioned speech head largely does not propagate it to output prosody, only a faint loudness echo survives. Realising prosody conditioning would require explicit prosodic supervision or the unified native-delivery model (which has the query audio in-context during generation), not the text-conditioned head as trained.

5.5 Training stability: attention-sink NaN and optimizer poisoning

Two instabilities cost real debugging time and generalize beyond this system. *(i) Attention-sink NaN.* The joint/grounded layouts place a raw audio embedding at sequence position 0. Transformer decoders calibrate position 0 as an attention sink (a massive-activation token, normally BOS); a raw audio embedding there makes the backbone's RMSNorm `rsqrt` backward singular, producing a *NaN gradient from a finite loss.* Global gradient-norm clipping then multiplies that single NaN into every parameter at the first optimizer step, which looks like sudden whole-model corruption. `torch.autograd.set_detect_anomaly` localizes it to RMSNorm; the fix is to prepend a BOS token so position 0 is calibrated, plus a finiteness guard that skips any step with a non-finite gradient norm. The transferable lesson: a finite loss does not imply finite gradients, and clipping operates after the NaN already exists. *(ii) Optimizer poisoning on resume.* `PagedAdamW8bit` quantizes the optimizer's second moment; a quantized near-zero variance makes the Adam update $m/(\sqrt{v}{+}\epsilon)$ explode *inside* the optimizer, downstream of gradient clipping, so clipping cannot prevent it (a

Table 5: Teacher-forced c_0 versus free-running WER, OpenS2S Phase-A 12B. Low c_0 coexists with garbled generation; only WER exposes it.

Model (training pairs)	c_0 trained	c_0 unseen	decoded WER
ckpt_6000 (636)	2.31	8.18	,
ckpt_14000 (2097)	1.56	2.09	141%

resumed run diverged to depth loss >200). Resuming with a fresh full-precision optimizer (which fits on the 48 GB rental, though not on 16 GB) removes it.

5.6 Key finding: c_0 is not intelligibility

An early 12B Phase-A run on a small conversational corpus (OpenS2S) drove teacher-forced c_0 to 1.3, which resembled our intelligibility breakthrough but was memorization. Table 5 shows the dissociation: held-out c_0 is far worse than trained c_0, and the decoded audio is unintelligible (141% WER, 12% recall) even where c_0 looks healthy.

This dissociation mirrors Penov [12]: there, aggregate language-modeling loss (bits-per-byte) cannot detect sensor conditioning, a capacity-matched constant-feature control reproduces the entire metric gain, and only task-aligned probes expose it. We observe the analogous gap between teacher-forced c_0 (low even when the model cannot speak) and free-running WER, which we therefore adopt as the gate. The same lesson recurs across the E2B series: aggregate depth cross-entropy repeatedly failed to predict perceptual quality, so word recall and blind listening, not loss curves, measure speech.

5.7 Codec entropy floor

The acoustic ("depth") loss plateaus at ∼6.7 and does not fall further. Scaling the depth transformer from 150M to ∼425M parameters did not lower the plateau, so it is not a capacity limit; it tracks Mimi's intrinsic per-codebook entropy (∼6.3–6.6). M5 confirmed this independently: a 30× increase in the depth-loss weight (0.01 → 0.3) produced *zero* reduction in held-out acoustic loss (validation depth flat at 6.55–6.60 across 2–10k steps), no intelligibility gain, and introduced stutter (repeated valid codec frames, invisible to per-frame cross-entropy). Cross-entropy over discrete RVQ codes therefore caps naturalness; closing it requires a flow-matching or diffusion acoustic head over continuous features, not a loss-weight knob.

6 Discussion & Limitations

Three limitations stand out. First, the speech head is trained from scratch at d=3840, so Phase A needs a large clean corpus; our OpenS2S misstep shows that too little data produces a memorizing model that looks good on c_0 but cannot speak. Second, naturalness is bounded by the codec floor (§5.7), not by model capacity or loss weighting; a continuous, flow-matching acoustic head is the indicated next step. Third, single-16 GB-GPU training is memory-bound: even a rented A100 gave only ∼1.5× throughput, so scaling requires configuration changes (sequence/precision/loader), not just a bigger card. Fourth, while the unified single model is demonstrated (§5.3), one native-delivery forward, intelligible and grounded full-length speech, its out-of-distribution competence is eroded by adapter training: on prompts outside the OpenS2S distribution it returns training templates, even though a three-way ablation shows the frozen backbone's reasoning is intact (native Gemma and

the adapter-disabled model both answer correctly; only the trained LoRA masks it). The decisive delivery insight still holds, a hand-built feature-splice carries prosody but not lexical content, while the backbone's own multimodal pathway carries both, and the remaining problem is preserving the backbone's latent reasoning while wiring the head onto it. Concurrent text-free systems chart the route: Qwen2.5-Omni [19] conditions its speech generator on hidden representations from a genuine native forward, and MOSS-Speech [22] carves speech-prediction layers into the trunk, frozen first, then partially unfrozen, on far more data, rather than appending an external head behind a heavy adapter. Both point the same way as our ablation: touch the "thinker" less (lighter or no LoRA, broader data, or in-trunk modality layers), so the frozen backbone's reasoning survives into the spoken reply.

7 Conclusion

We show that lightweight adapters can teach a fully frozen, quantized LLM to speak intelligibly (13% WER), and that the same frozen, encoder-free backbone can hold a spoken conversation through a two-stage cascade (native-pathway understanding into the backbone, then a CSM-style speech head) with no backbone training. Along the way we document five transferable findings. First, naive 4-bit quantization silently destroys a multimodal backbone's audio comprehension while sparing prosody. Second, *how* audio is delivered (native pathway vs. hand-built feature-splice), not quantization or capacity, decides whether lexical content survives. Third, a convenient teacher-forced loss is not a proxy for the capability of interest, for both speech and comprehension (WER on decoded audio is the gate). Fourth, a discrete neural codec imposes an entropy floor that cross-entropy cannot cross. Fifth, when the cascade is collapsed into a single native-delivery model, the frozen backbone's reasoning stays intact but heavy adapter training masks it out-of-distribution: a three-way ablation restores correct replies by disabling the adapter, so robustness is a data-and-adaptation problem, not an architectural one. That single end-to-end model, with the speech head conditioned on the backbone's own latent reply, speaks intelligible, grounded, full-length replies in one forward pass, realizing the core thesis of the model's own latent reasoning spoken aloud.

Reproducibility

Code: `https://github.com/giannisanni/encoder-free-speech`. Checkpoints and data shards are mirrored to HuggingFace (`giannisan/gemma-csm-ckpts`, `giannisan/gemma-csm-shards`). The appendix records the exact environment and commands.

References

[1] Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2204.14198.

[2] CASIA-LM. OpenS2S: Advancing fully open-source end-to-end empathetic large speech language model. *arXiv preprint arXiv:2507.05177*, 2025. Apache-2.0. EN split: 49,368 paired (audio-query, audio-response) examples w/ transcripts + emotion; response wavs in 7 tar parts. Native tokenizer is GLM-4-Voice; we re-encode response audio with Mimi.

[3] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[4] 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. Introduces the Mimi codec: streaming split-RVQ, 24 kHz, 12.5 Hz frame rate, codebook 0 semantic + acoustic; checkpoint supports up to 32 codebooks (Moshi operates at 8 @ 1.1 kbps).

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

[6] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.06666.

[7] Google DeepMind. Gemma 4 12b: a unified, encoder-free multimodal model. Google technical report / model card. `https://ai.google.dev/gemma/docs/core/model_card_4`, `https://huggingface.co/google/gemma-4-12B`, 2026. No arXiv report. 12B "Unified" is encoder-free: 16 kHz audio in 40 ms / 640-float frames linearly projected to hidden 3840; 48 layers. NB: E2B/E4B audio variants use a USM conformer encoder (NOT encoder-free).

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1902.00751.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

[10] Kushal Lakhotia et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1336–1354, 2021. arXiv:2102.01192.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.08485.

[12] Franci Penov. Conditional lora bridges for modular sensor adaptation of frozen small language models. kortexa.ai (LegoLM), self-published. `https://github.com/kortexa-ai/legolm.paper`, June 2026. Hypernetwork generates LoRA from frozen sensor-encoder features for a frozen 33.6M LM. Key finding (parallels ours): aggregate LM loss (BPB) cannot detect conditioning; task-aligned probes can.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2212.04356.

[14] Paul K. Rubenstein et al. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

[15] Johan Schalkwyk, Ankit Kumar, Dan Lyth, Sefik Emre Eskimez, Zack Hodari, Cinjon Resnick, Ramon Sanabria, and Raven Jiang. Crossing the uncanny valley of conversational voice (CSM). Sesame AI research. `https://www.sesame.com/research/crossing_the_uncanny_valley_of_voice`; code `https://github.com/SesameAILabs/csm`, 2025. No arXiv paper. Backbone

(Llama) predicts Mimi codebook 0 (semantic); a depth decoder predicts the remaining 31 acoustic codebooks. We adopt this head topology on a frozen Gemma-4 backbone.

[16] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.13289.

[17] Chengyi Wang et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[18] Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2411.00774. Frozen LLM; streaming speech encoder + adapter on input; AR single-codebook speech decoder driven by the LLM output; freezing avoids catastrophic forgetting under limited data.

[19] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. Thinker-Talker: the Talker "directly utilizes the hidden representations from the Thinker to produce audio tokens," trained end-to-end (backbone not frozen).

[20] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech*, pages 1526–1530, 2019. arXiv:1904.02882. 585 h, 24 kHz, 2456 speakers, CC BY 4.0. We use a 110 h subset.

[21] 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. In *Findings of EMNLP*, 2023. arXiv:2305.11000.

[22] Xiaoran Zhao, Ben Wang, et al. Moss-speech: Towards true speech-to-speech models without text guidance. *arXiv preprint arXiv:2510.00499*, 2025. Qwen3-8B, 36 layers; first 32 shared trunk, last 4 split into parallel text/speech 4-layer stacks (modality-based layer-splitting). Stage 1 freezes the backbone and trains only speech embeddings, speech layers, and speech head; stage 2 unfreezes. Text-free S2S; single-codebook 12.5 Hz tokenizer (ASR-objective) + flow-matching decoder.

A Reproducibility build-log

Backbone. `google/gemma-4-12B-it`, `model_type gemma4_unified`; `AutoModelForCausalLM` returns `Gemma4UnifiedForConditionalGeneration`, whose `.model` is `Gemma4UnifiedModel` (owns `get_audio_features`, `embed_audio`, `language_model`). Text config: hidden 3840, 48 layers, GQA 8 KV, no per-layer-embedding tables, no KV-sharing. Audio config: hidden 640, `audio_embed_dim` 640, `audio_token_id` 258881, up to 30 s of audio per input.

Environment. A dedicated venv (`audioin-venv`): torch 2.11.0+cu128, torchvision, transformers built from source (5.13.0.dev0, which ships `Gemma4UnifiedProcessor` / `Gemma4UnifiedAudio FeatureExtractor`), bitsandbytes, peft, pillow, soundfile, accelerate, datasets. The m-series used a separate venv with transformers 5.10.2.

Data pipeline. LJSpeech and LibriTTS are resampled to 24 kHz mono and Mimi-encoded (all 32 codebooks) into `{id, text, codes}` shards (∼18 clips/s on the consumer card). OpenS2S English (49,368 pairs; response audio ∼71 GB across `en_response_wav.tar.gz.0-6`) is too large to download; we stream first-encountered wavs from each tar (the tars are not index-ordered) and intersect query/response/manifest indices, then re-encode with Mimi. A chained multi-piece reader streams `.0`–`.6` as one continuous gzip; a fresh `HfFileSystem` is required for the query pass (closing the response handles tears down the shared HTTP client).

Training stability. Checkpoints are written atomically (`.tmp` then `os.replace`); a non-atomic save interrupted by a process kill truncated a checkpoint and broke resume. `-resume` restores model, step, and optimizer. A supervisor auto-resumes on death and frees VRAM (headless target; stop GPU-resident side services) before each launch. Free-running decode is gated with a silence-retry, and the run-log is appended (not truncated) across restarts so the TensorBoard tail-follower stays continuous.