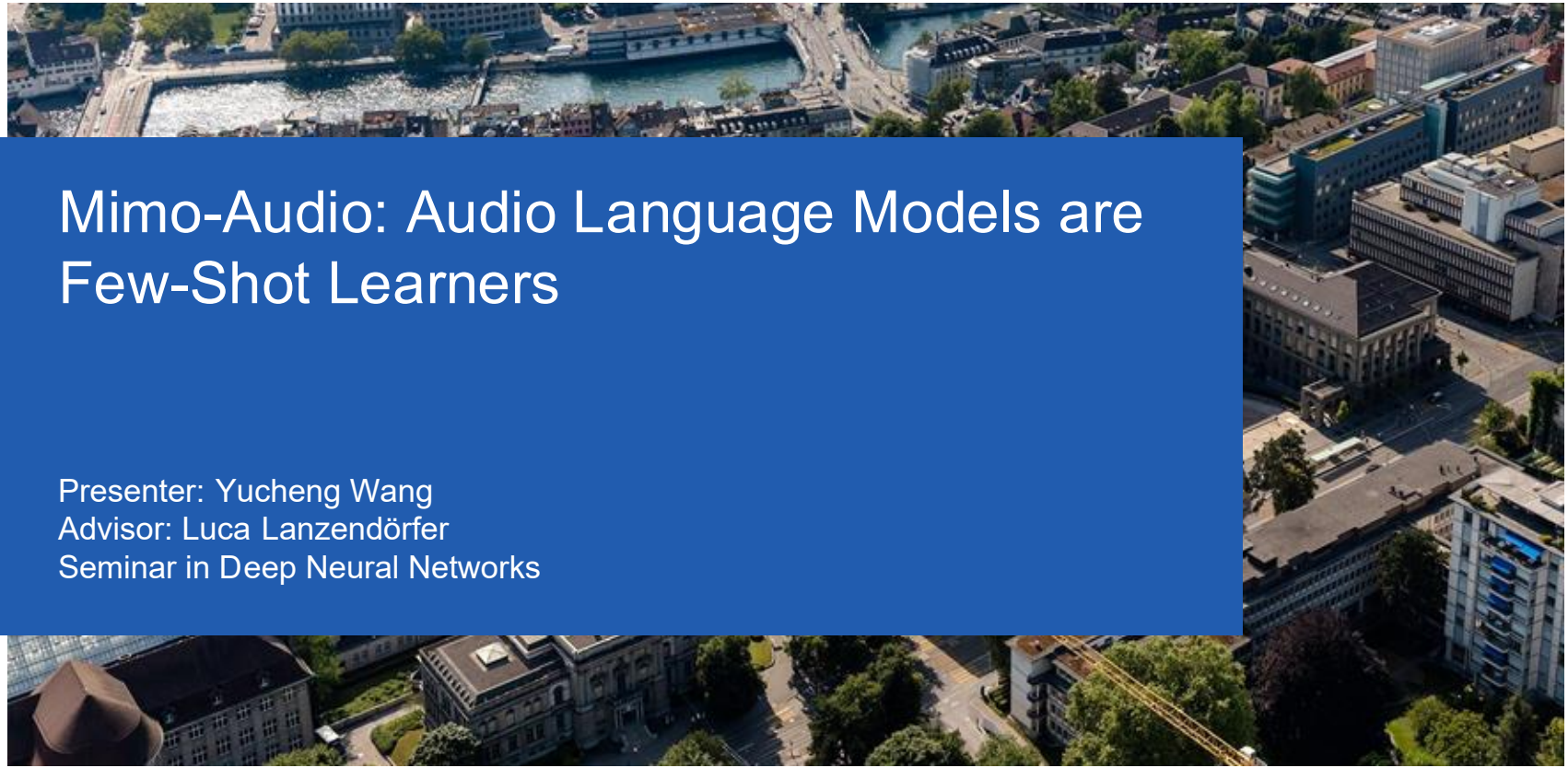
ETH zürich



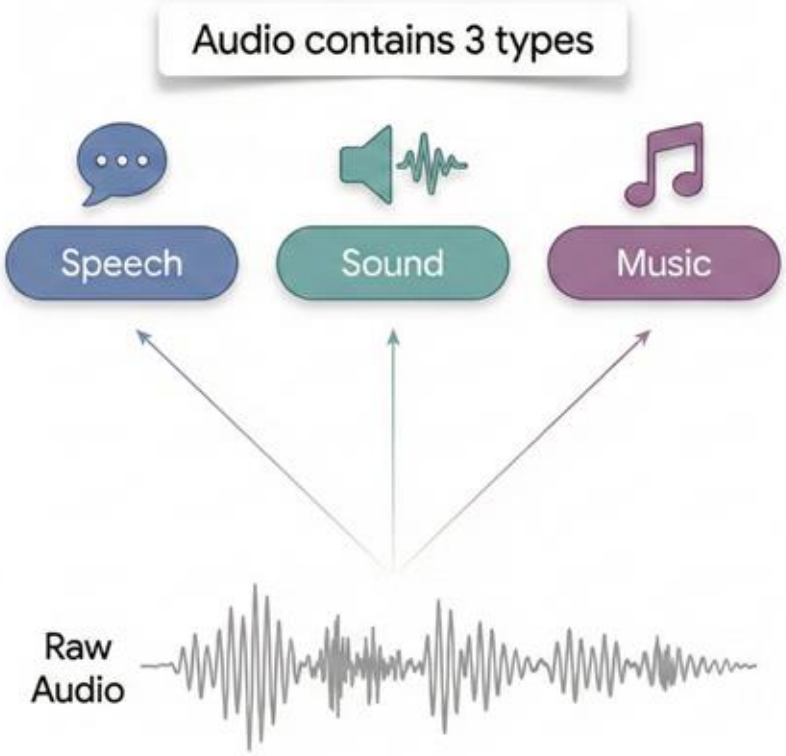
Mimo-Audio: Audio Language Models are Few-Shot Learners

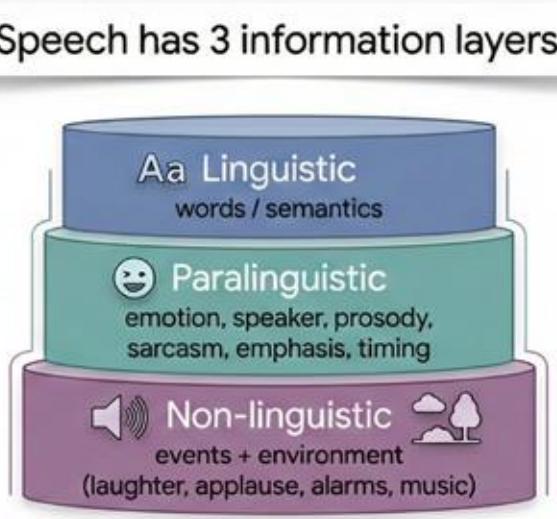
Presenter: Yucheng Wang
Advisor: Luca Lanzendörfer
Seminar in Deep Neural Networks

Mimo-Audio is an Audio Large Language model that can understand and generate audios.

But why do we need audio understanding?

Speech/Audio Understanding ≠ Speech Transcription

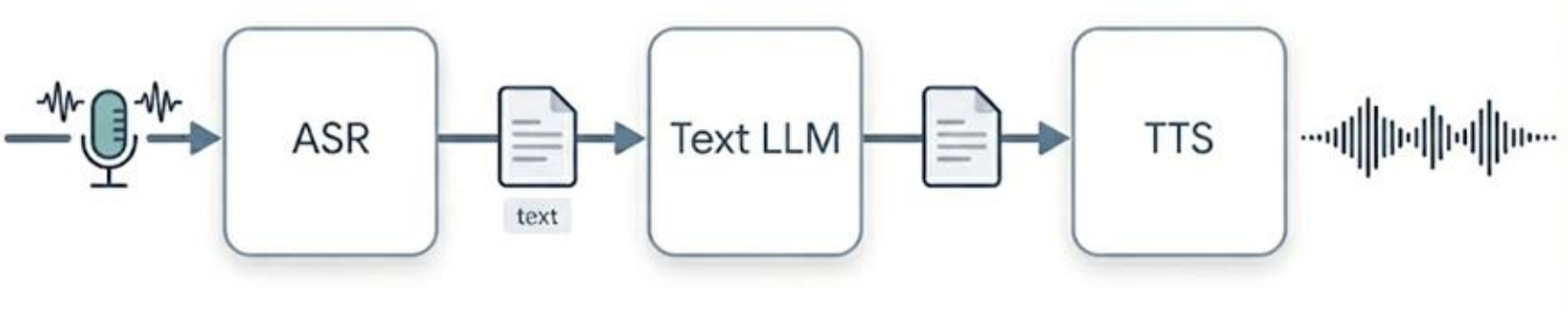




ASR(Speech transcription) captures mainly **Linguistic** content — but meaning also lives in **Paralinguistic** + **Non-linguistic** cues.

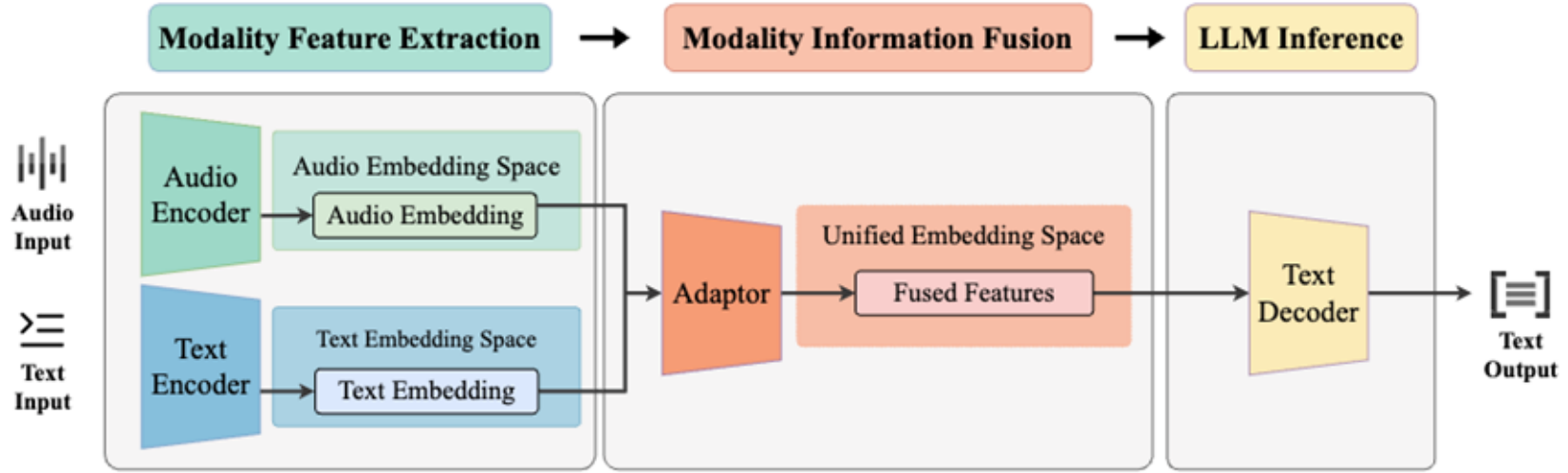
Why do we need End-to-end Large Audio Language Models?

Traditional cascaded system:



We want a **unified representation** that preserves linguistic + paralinguistic + non-linguistic cues

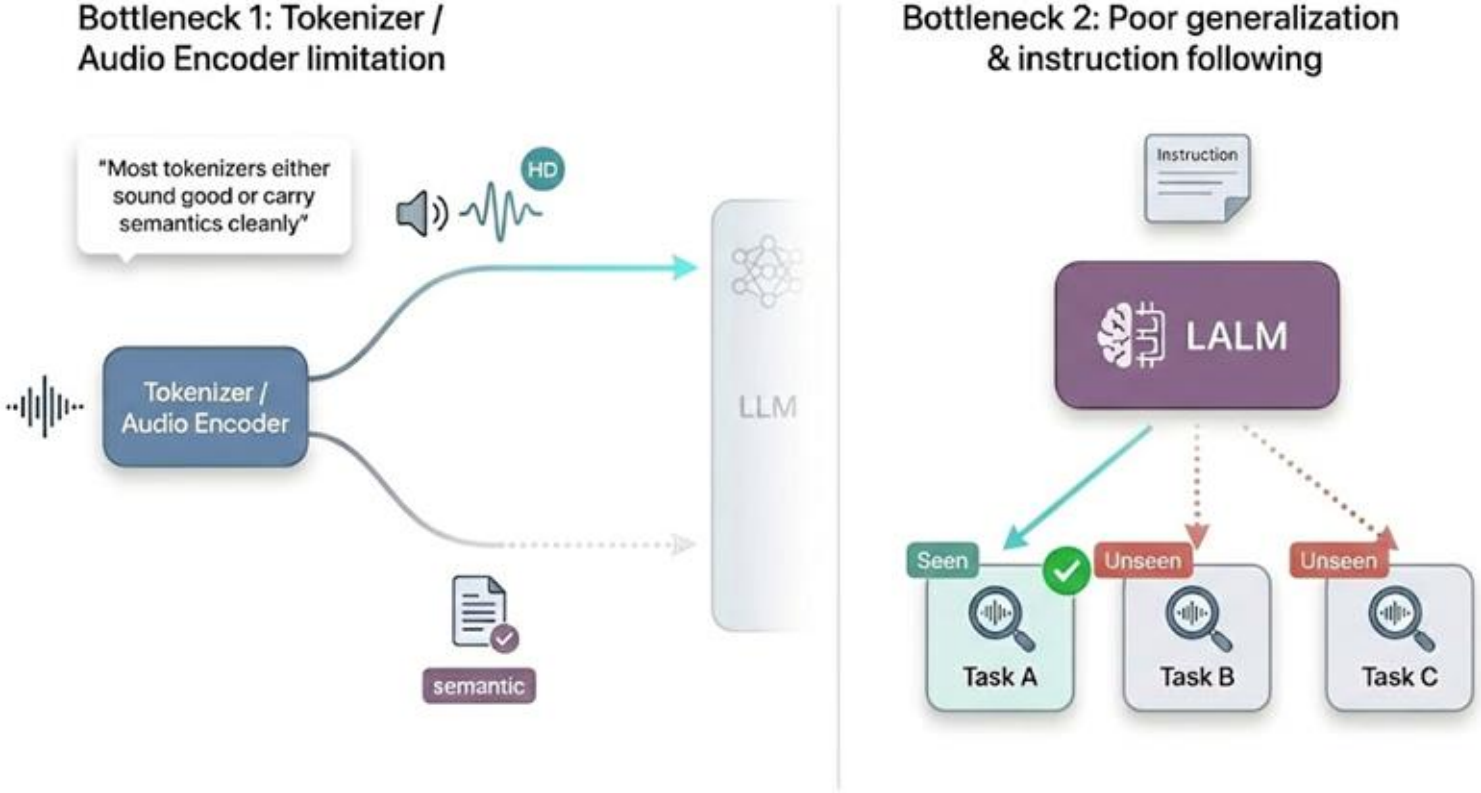
General Architecture & Training of an LALM



A Survey on Speech Large Language Models for Understanding [Peng, etc. 2025]

Training:
Modality Alignment + Multi-task Supervised Finetuning + Post-training (very limited)

Bottlenecks of current LALMs



Mimo-Audio's Bet

BOTTLENECK 1 – REPRESENTATION

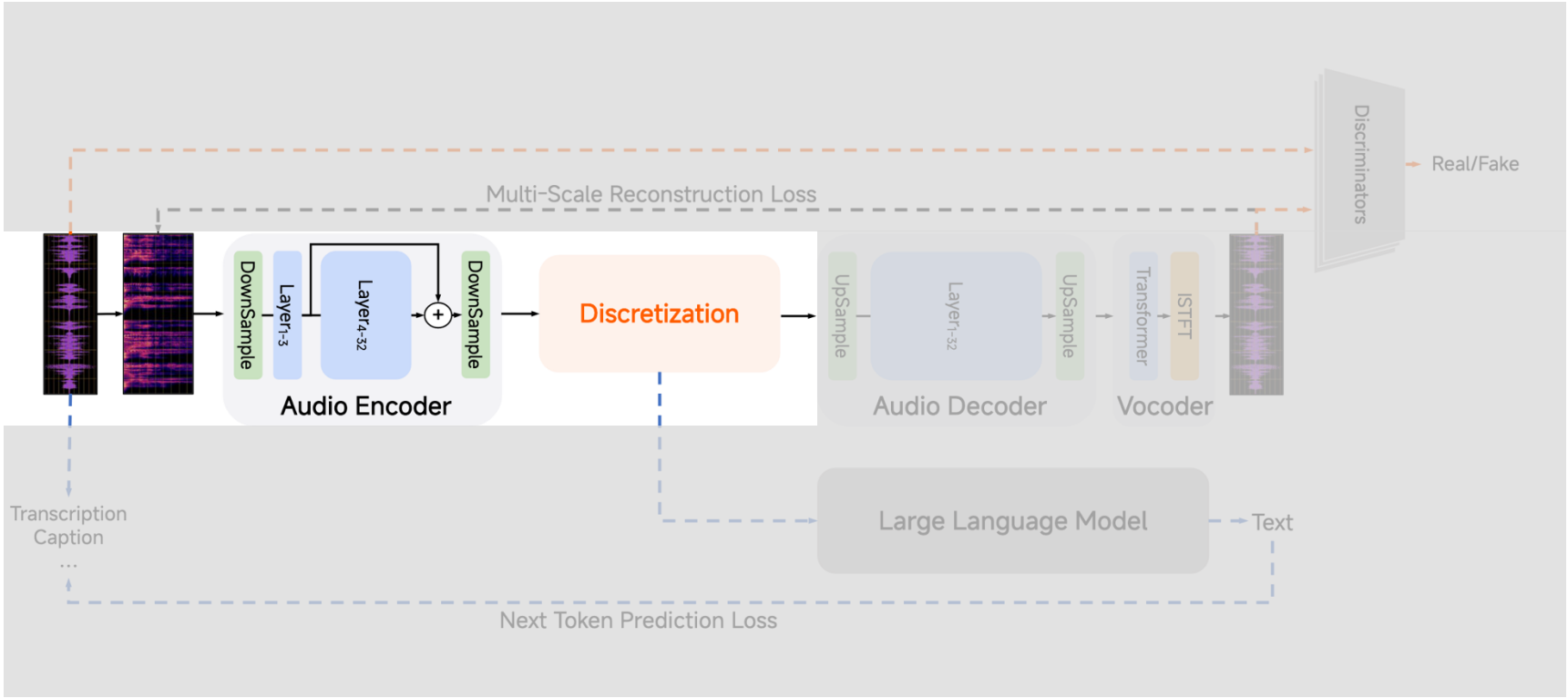
Bet 1 – A unified tokenizer can preserve both acoustic detail and linguistic semantics.

BOTTLENECK 2 – GENERALIZATION

Bet 2 – Scale + next-token training can unlock task-agnostic few-shot behavior (à la GPT-3)

Idea of GPT 3: scale → emergent few-shot 'in-context' learning.

MiMo-Audio-Tokenizer

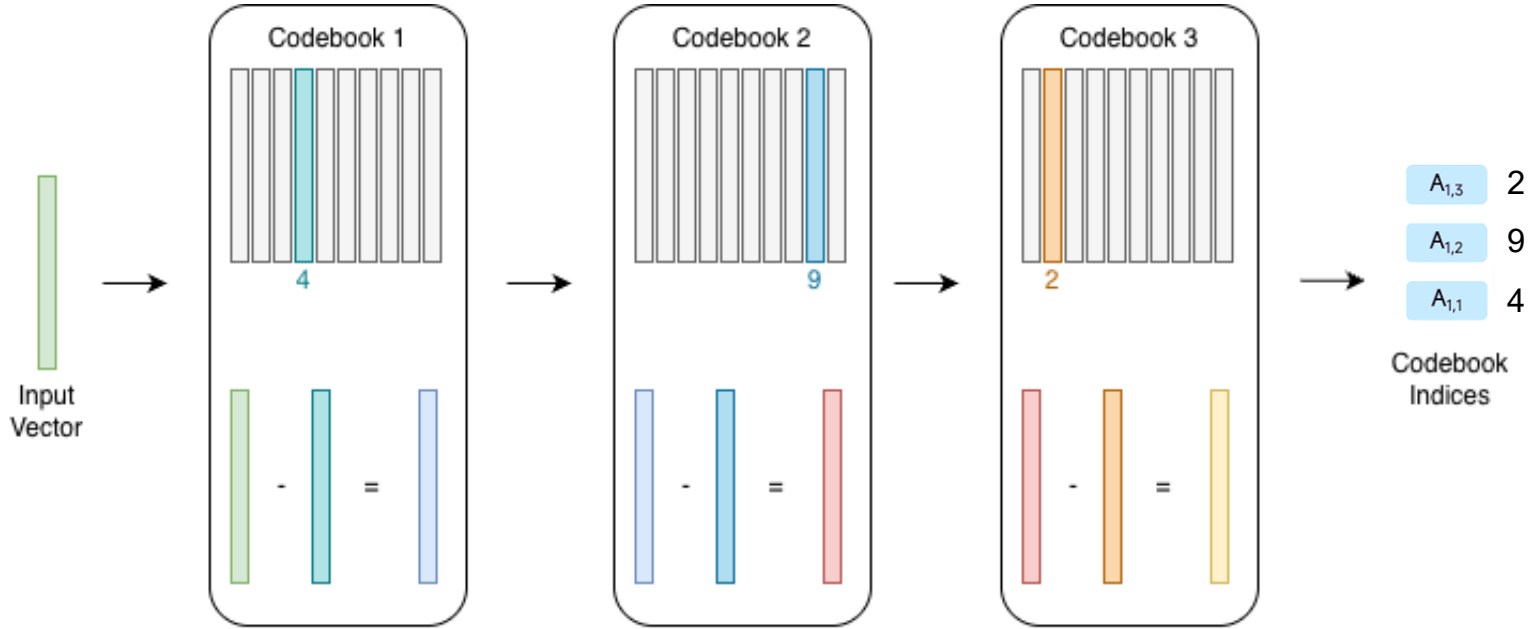


MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

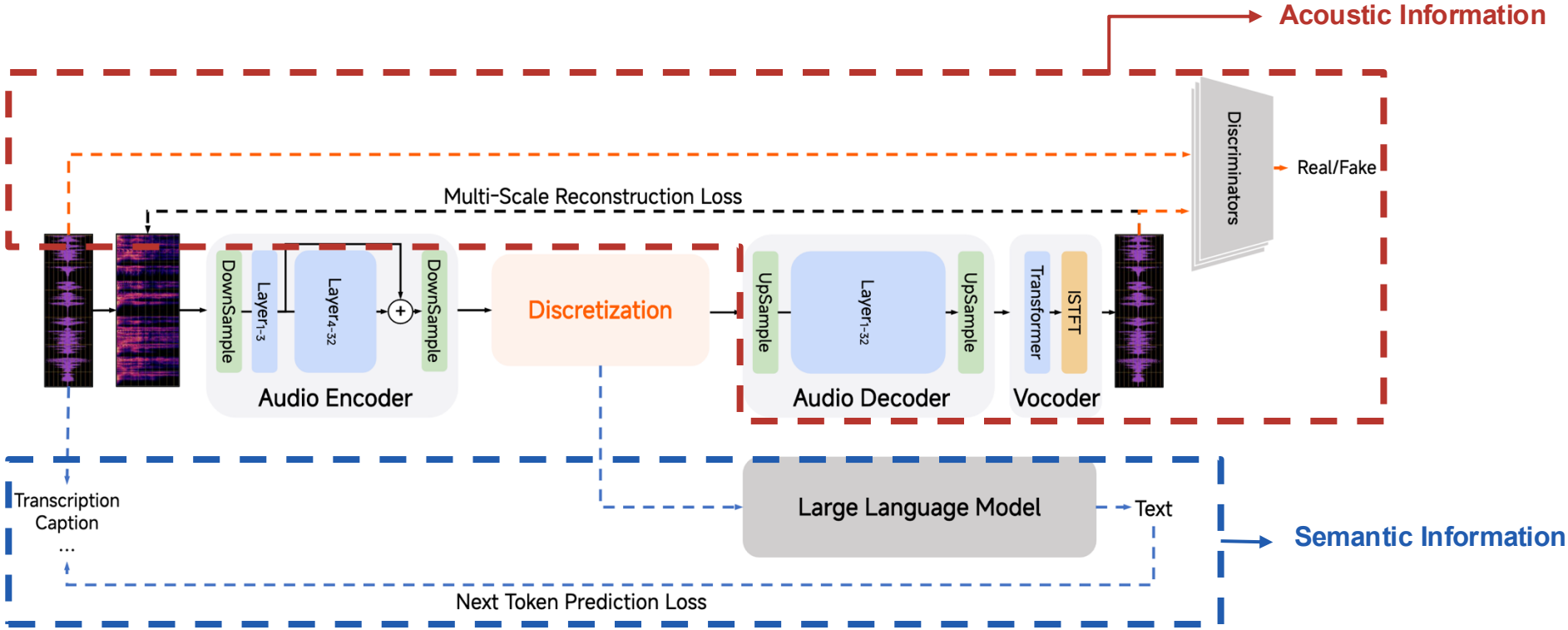
Background: RVQ

Residual Vector Quantization (RVQ) is a way to turn a continuous audio feature into discrete tokens **in multiple steps**.

RVQ token (The output of RVQ) for one frame is actually a **stack of code indices [k_1, k_2, k_3, …]**



MiMo-Audio-Tokenizer



MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

When mixing text + audio tokens in one context...

Information density mismatch

Common Frame rate: 12.5–50 Hz (20–80 ms/token)

Text tokens (prompt)

mismatch

Audio tokens (speech)

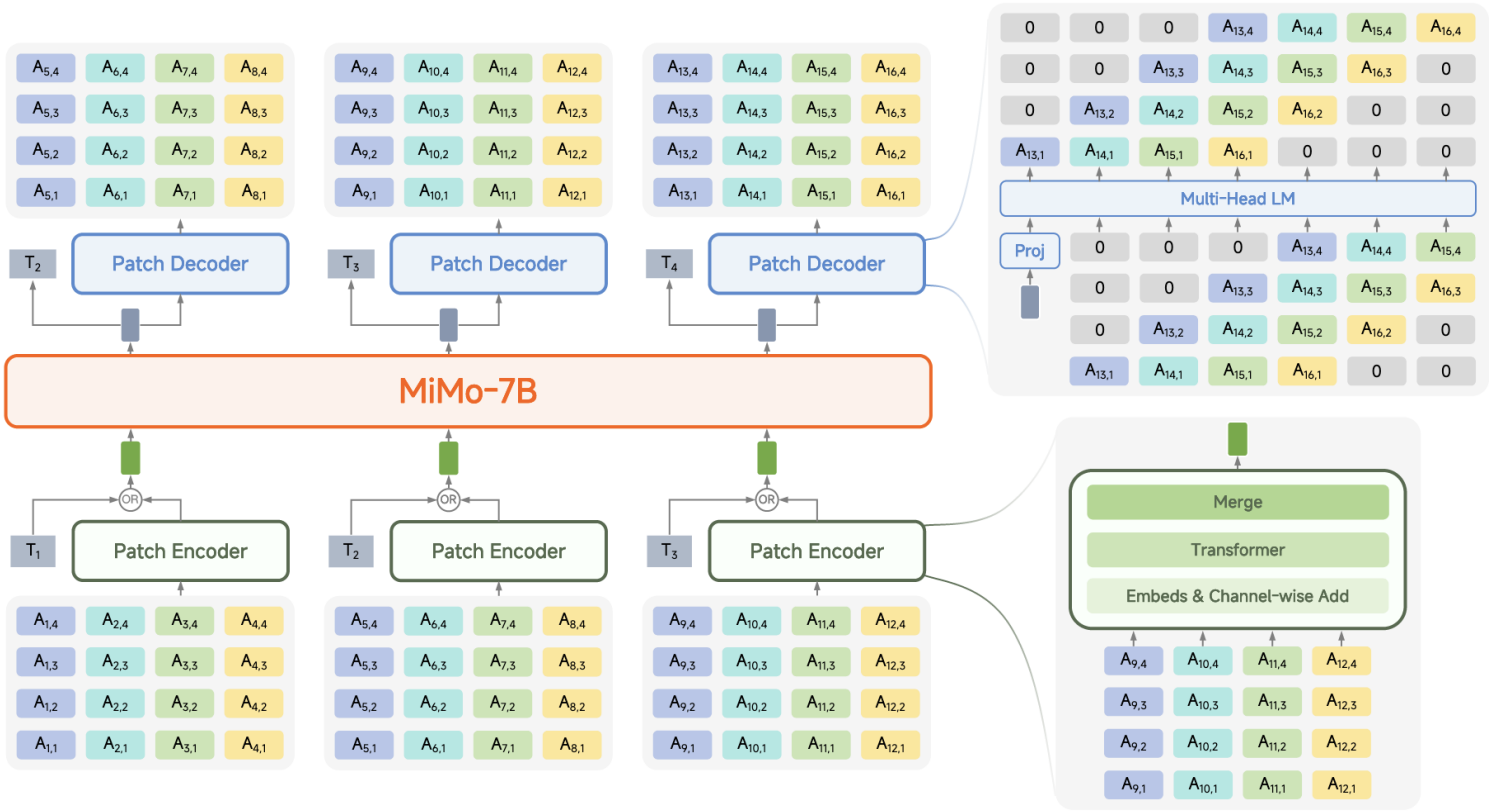
Quick intuition

- 4s speech → ~50–200 audio tokens (at 12.5–50 Hz)
- Text prompt: far fewer tokens can express comparable meaning

Two design choices to reduce mismatch

- Lower token rate: 6.25 Hz
- Patch design: group G frames → one audio patch token

MiMo-Audio Model Architecture

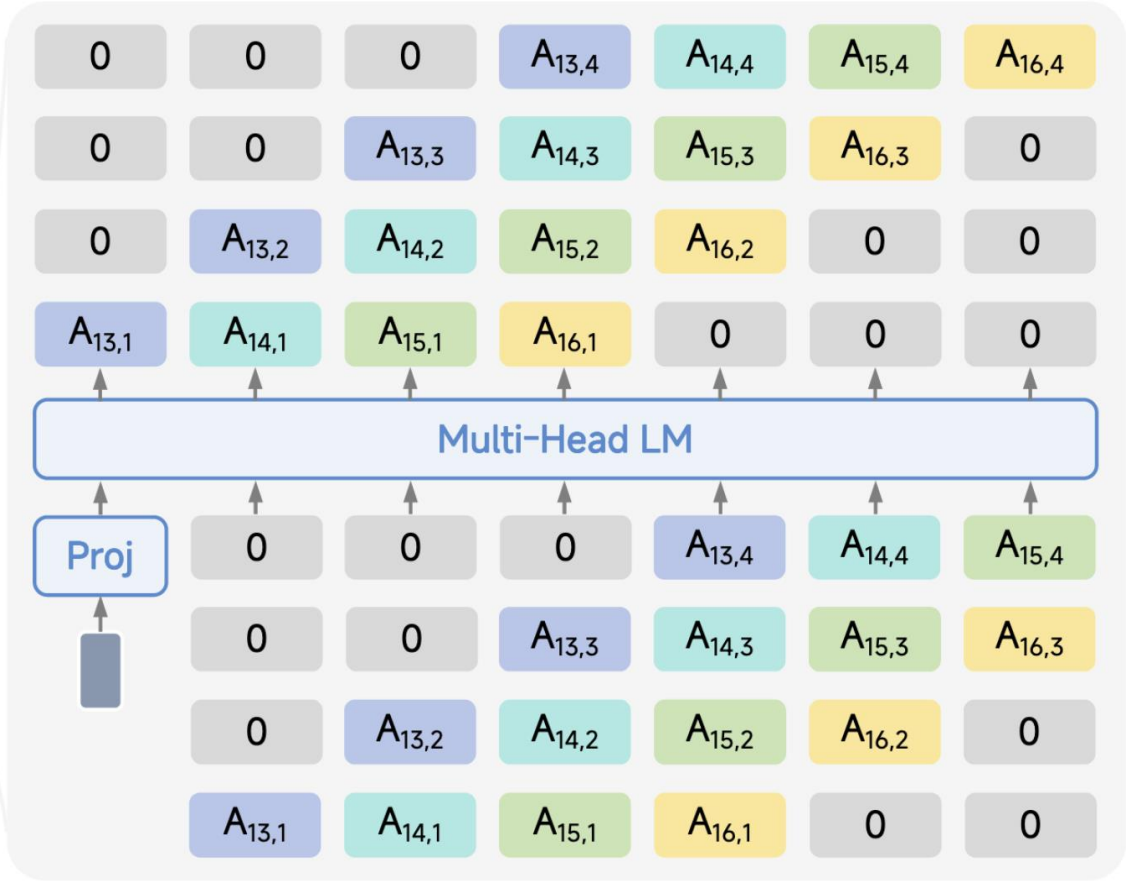


MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

Delayed Generation

RVQ codebooks are **hierarchical**: finer codes depend on coarser codes **at the same frame**.

Solution: shift codebooks in time



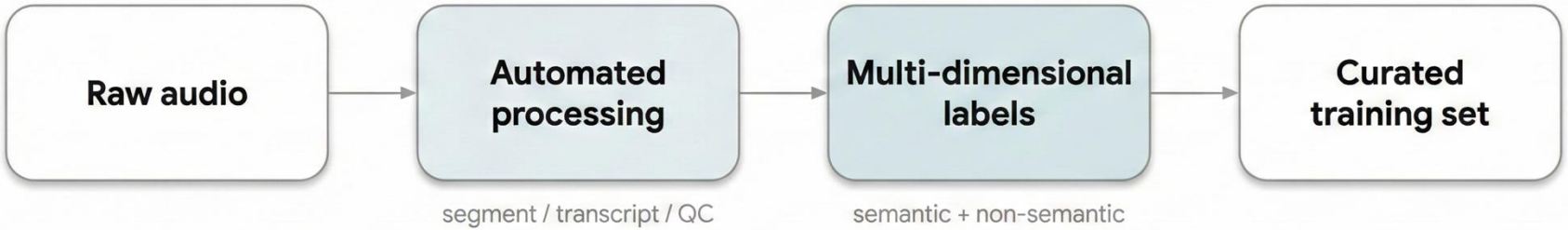
MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

Pretraining Data: Scale + Automatic Curation

Text-only + Speech-only + Speech–Text pairs

Hundreds of millions of hours of in-the-wild speech (diverse)

Automated pipeline + multi-dimensional labels → filtering & sampling



Two-stage Pretraining

Stage 1 – Understanding Training

(A) Training Components

(B) Training Tasks

- Speech-Text interleaved
- ASR
- Audio captioning
- Text-only pretraining

Stage 2 – Understanding-Generation Joint Training

(A) Training Components

(B) Training Tasks

- Speech continuation • ASR • TTS
- Speech-Text interleaved
- Audio captioning • Instruction-following TTS
- Text-only pretraining

Pretraining Evaluation

1. Few-shot In-context Learning

Capabilities

- General Knowledge
- Audio Understanding
- Speech-to-Speech

→

K-shot

Few-shot prompt (K examples)

Example 1: input → output

Example 2: input → output

Query: input → ?

→

MiMo-Audio

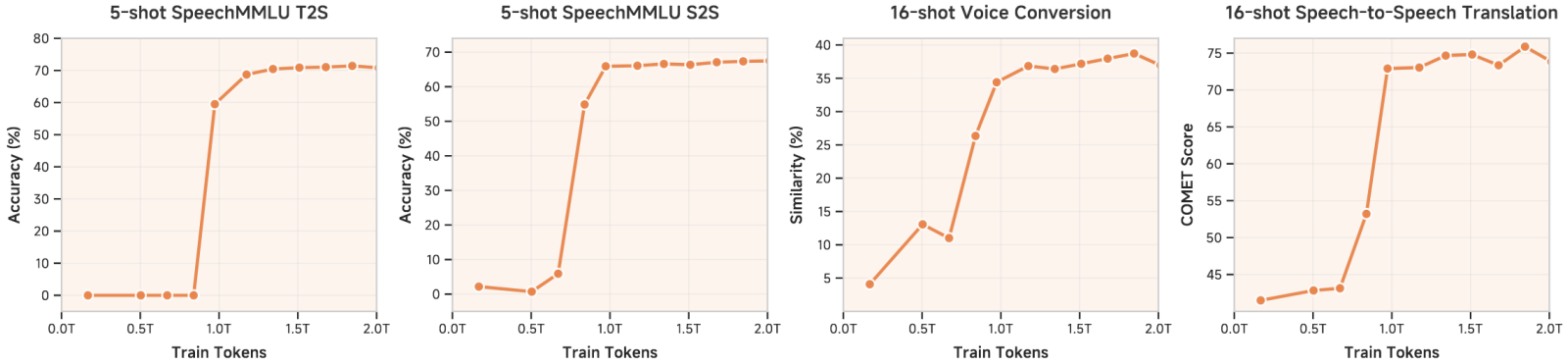
Infers the task from the prompt (no fine-tuning).

2. Speech Continuation

Evaluated on its ability to seamlessly continue a given speech prompt, maintaining speaker identity and prosody.

Pretraining Results

- **Emergent Ability**



MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

- **Demo**

Post-training

Overview & Data Types

Activate pre-trained model's **understanding & generation** across tasks via supervised instruction finetuning datasets.

3 Major Post-training Data Types

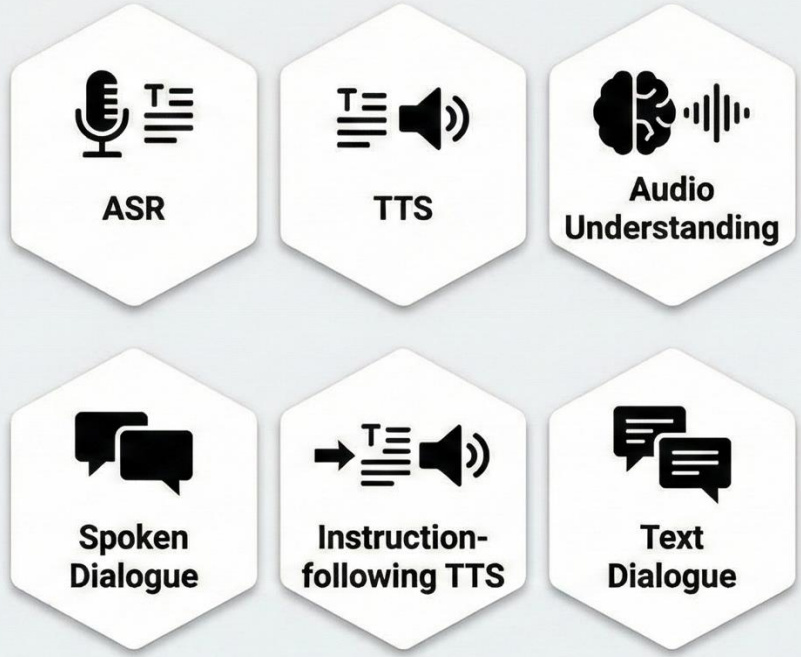
1. Audio Understanding

- Audio captioning
- Audio question answering

2. Speech Generation (Instruction following)

3. Spoken Dialogue

6 Task Types in Post-training



Post-training Evaluation

Datasets	Model	Performance
	Audio Understanding	
MMAU Speech | Sound | Music | Overall	MiMo-Audio-7B-Instruct	68.47 | **82.58** | **73.65** | **74.90**
	Gemini 2.5 Flash	**76.58** | 73.27 | 65.57 | 71.80
	Audio Flamingo 3	66.37 | 79.58 | 66.77 | 73.30
	Step-Audio2-mini	68.16 | 79.30 | 68.44 | 72.73
	Kimi-Audio-Instruct	62.16 | 75.68 | 66.77 | 68.20
	Qwen2.5-Omni	**70.60** | 78.10 | 65.90 | 71.50
	GLM-4-Voice	35.44 | 27.63 | 27.84 | 30.30
MMAU-Pro	MiMo-Audio-7B-Instruct	**53.35**
	Gemini 2.5 Flash	**59.20**
	Audio Flamingo 3	51.70
	Step-Audio2-mini	47.91
	Kimi-Audio-Instruct	46.60
	Qwen2.5-Omni	52.20
	GLM-4-Voice	38.25
	GPT-4o-Audio	52.50
MMAR	MiMo-Audio-7B-Instruct	**63.60**
	Gemini 2.5 Flash	**65.60**
	Audio Flamingo 3	58.50
	Step-Audio2-mini	55.80
	Kimi-Audio-Instruct	48.00
	Qwen2.5-Omni	56.70
	GLM-4-Voice	29.50
	GPT-4o-Audio	63.50
MMSU Perception | Reasoning | Overall	MiMo-Audio-7B-Instruct	**46.86** | 76.98 | **61.70**
	MiMo-Audio-7B-Instruct +Think	**51.71** | 74.79 | **62.88**
	Gemini 1.5 Pro	- | - | 60.70
	Audio Flamingo 3	- | - | 61.40
	Step-Audio2-mini	42.71 | 72.60 | 57.18
	Kimi-Audio-Instruct	44.84 | 75.70 | 59.78
	Qwen2.5-Omni	42.67 | **77.64** | 58.10
	GLM-4-Voice	11.04 | 16.16 | 13.30

	Spoken Dialogue	
Big Bench Audio S2T | S2S	MiMo-Audio-7B-Instruct	**72.90** | **60.20**
	gpt-4o-audio-preview-2024-12-17	70.20 | **67.20**
	Step-Audio2-mini	50.90 | 47.50
	Kimi-Audio-Instruct	59.40 | 51.00
	Qwen2.5-Omni	54.20 | 53.60
	GLM-4-Voice	44.80 | 42.70
MultiChallenge Audio S2T | S2S	MiMo-Audio-7B-Instruct	**15.15** | **10.10**
	Step-Audio2-mini	13.64 | 8.08
	Kimi-Audio-Instruct	7.07 | 1.01
	Qwen2.5-Omni	11.11 | 8.08
	GLM-4-Voice	9.09 | 6.06

Datasets	Model	Performance
	TTS	
Seed-TTS-Eval ZH | EN | ZH-Hard	MiMo-Audio-7B-Instruct	**1.96** | 5.37 | **14.14**
	Step-Audio2-mini	2.13 | **3.18** | 16.31
	Instruct-TTS	
InstructTTSEval-EN APS | DSD | RP | Overall	MiMo-Audio-7B-Instruct	**80.60** | **77.63** | **59.54** | **72.59**
	GPT-4o-mini-tts	76.40 | 74.30 | 54.80 | 68.50
InstructTTSEval-ZH APS | DSD | RP | Overall	MiMo-Audio-7B-Instruct	**75.74** | **74.3** | **61.54** | **70.52**
	GPT-4o-mini-tts	54.90 | 52.30 | 46.0 | 51.07
	ASR	
ASR Librispeech-test-clean | AISHELL	MiMo-Audio-7B-Instruct	3.50 | 1.65
	Step-Audio2-mini	**1.87** | 0.95
	Kimi-Audio-Instruct	2.13 | **0.62**

MiMo-Audio: Audio Language Models are Few-Shot Learners [LLM-Core Xiaomi 2025]

Post-training Evaluation – Audio Reasoning

Prompt: Who is talking?

thinking...

1. **Analyze the User's Request:** The user wants to identify the person speaking in the audio clip.
2. **Listen to the Audio and Extract Clues:**
   - **Voice:** The speaker is an older male with a distinct, somewhat raspy voice and a recognizable speaking cadence.
   - **Content:** The speaker mentions several key things:
     - "I gave my best to you."
     - "It's been the honor of my lifetime to serve as your president." This is a direct statement of having been a U.S. president.
     - "I love the job, but I love my country more." This is a common sentiment expressed by presidents.
     - "All this talk about how I'm angry at all those people who said I should step down." This is a very specific and crucial clue. It refers to the period in late 2022 when then-U.S. President Joe Biden faced significant pressure from within his own party to withdraw from the presidential race.
     - "I've made a lot of mistakes in my career." This is a common political trope, but the context makes it specific to him.
     - "We're the United States of America!" - A patriotic closing line often used by presidents.
     - "God bless you all, and may God protect our troops." - A standard presidential closing remark.

Acoustic: Older male with raspy voice

3. **Synthesize the Clues:**
   - The speaker is a former U.S. president.
   - The reference to"people who said I should step down" directly points to the late 2022 debate and the subsequent calls for him to exit the presidential race.
   - The person who was the U.S. president at that time and faced those specific calls was Joe Biden.
   - The voice is a perfect match for Joe Biden's known speaking voice.
4. **Form a Conclusion:** Based on the content (specifically the reference to stepping down) and the voice, the speaker is Joe Biden.
5. **Construct the Final Answer:**
   - State the identity of the speaker clearly.
   - Provide the evidence from the audio that led to this conclusion, focusing on the most definitive clue (the"step down" comment).
   - Mention the supporting evidence, such as his reference to being president and the characteristic voice.
   - Add context, noting that this speech was given after he announced he was ending his re-election campaign.

Semantic: content often used by presidents

Based on the content and voice in the audio, the person speaking is **Joe Biden**.

Several key pieces of evidence from the clip confirm his identity:

- **Reference to being President:** He states,"It's been the honor of my lifetime to serve as your president."
- **The"Step Down" Comment:** The most definitive clue is when he says,"...all those people who said I should step down." This directly refers to the widespread calls for him to end his re-election campaign in the summer of 2024.
- **Voice and Cadence:** The speaker's voice, tone, and speaking style are unmistakably those of Joe Biden.
- **Closing Remarks:** He ends with the standard presidential closing,"God bless you all, and may God protect our troops."

This speech was given by President Biden after he announced he was ending his campaign for re-election but before he announced his endorsement of Vice President Kamala Harris.

Observations from hands-on testing

- **ASR/TTS quality** is noticeably below **specialized** ASR or TTS systems.
- **Instruction following** is still inconsistent; the model occasionally produces responses that **drift off-prompt**.
- The model supports a **“thinking” mode**, suggesting it was trained with **CoT-style data**, but the paper provides limited detail on this component. Overall, **reliable audio-grounded reasoning** remains an open challenge.

Key Takeaways (MiMo-Audio)

- **Audio can be few-shot:** scale + next-token training unlocks in-context behavior.
- **Representation matters:** Tokenizer + patches keep linguistic and acoustic cue.
- **Not solved yet:** instruction following and robustness still lag specialists.

Q&A