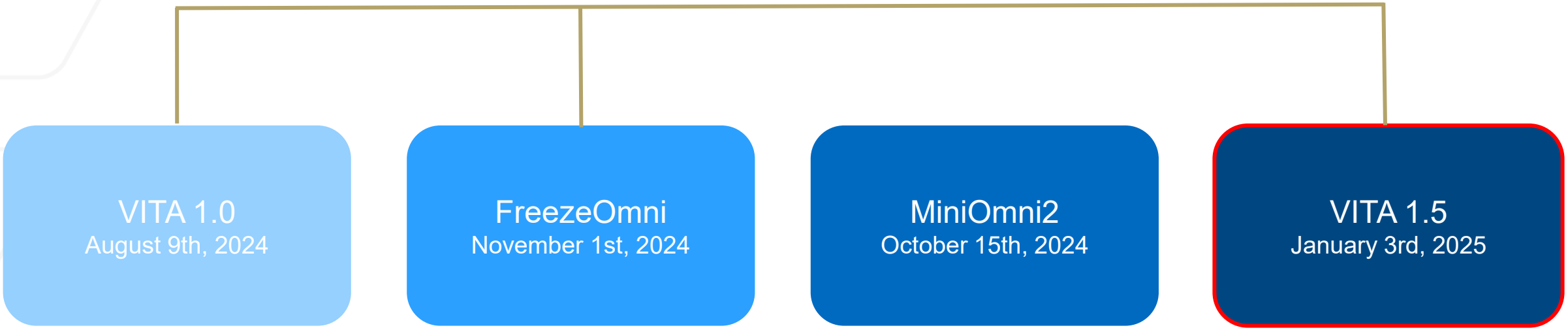
VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction

Seok Joon Kim, Chengyin Xu, Junghwan Yim

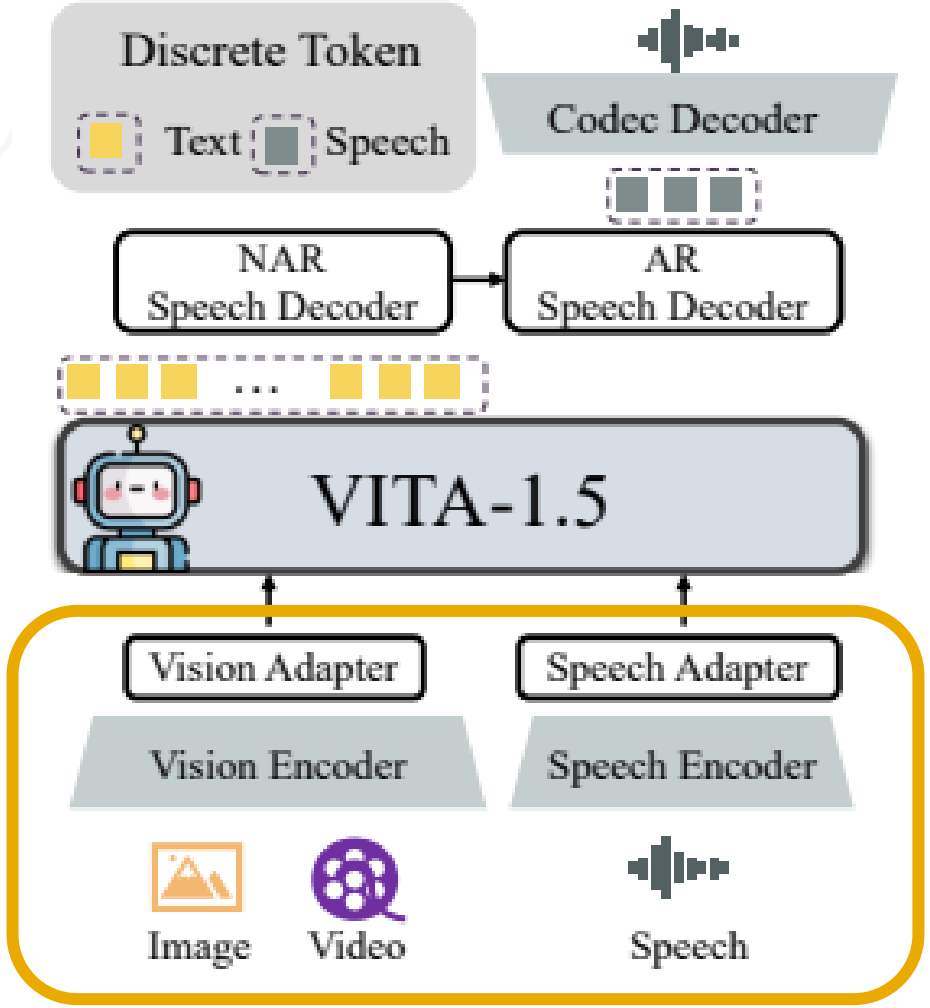
Georgia Tech

History



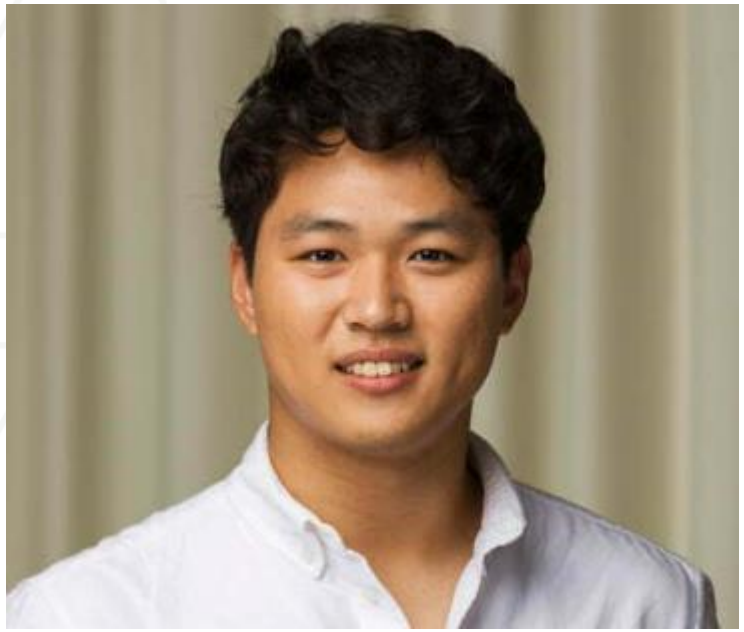
Fu et al, VITA: Towards Open-Source Interactive Omni Multimodal LLM https://arxiv.org/html/2408.05211v1
Wang et al, Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM https://arxiv.org/html/2411.00774v1
Xie et al, Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities https://arxiv.org/abs/2410.11190
Fu et al, VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction https://arxiv.org/abs/2501.01957

Architecture: Vision and Speech Encoders



- **Vision Encoder (InternViT-300M):** processes 448×448 pixel images into 256 visual tokens; dynamic patching preserves fine details in high-res images.
- **Video Processing:** uniform frame sampling (4–16 frames)
- **Vision Adapter (2-layer MLP):** projects high-dimensional visual features into LLM token space
- **Speech Encoder (350 M params):** 4× CNN downsampling + 24 Transformer layers (1024 hidden size); uses Mel-filterbank inputs (25 ms window, 10 ms shift).
- **Speech Adapter:** extra 2× CNN downsampling to match dimension with the LLM.
- Together, these modules form a **shared representation space** for vision, text, and audio

Member Introduction



Seok Joon Kim
1st year Robotics PhD (Mech)
seokjoonkim@gatech.edu

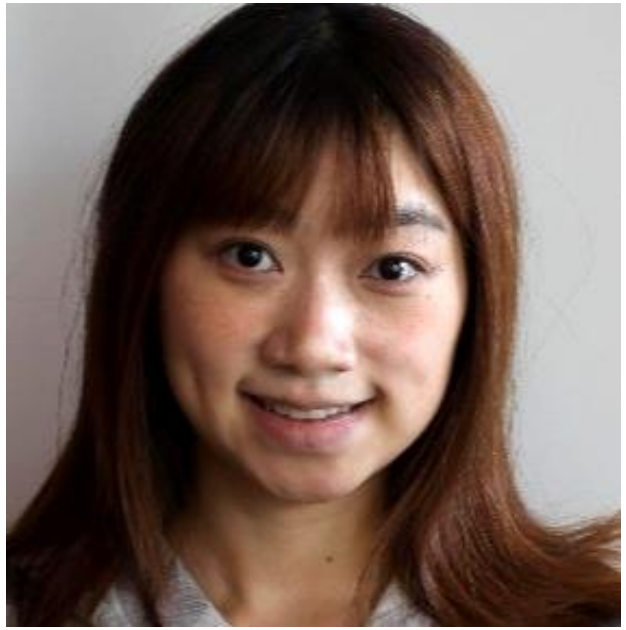
Symbiotic and
Augmented Intelligence Lab

Shared Autonomy, HCI / HRI

Junghwan Yim
1st year Mechanical Engineering PhD
jyim67@gatech.edu

SK Lab

Digital Twin, Physical AI, Agent AI



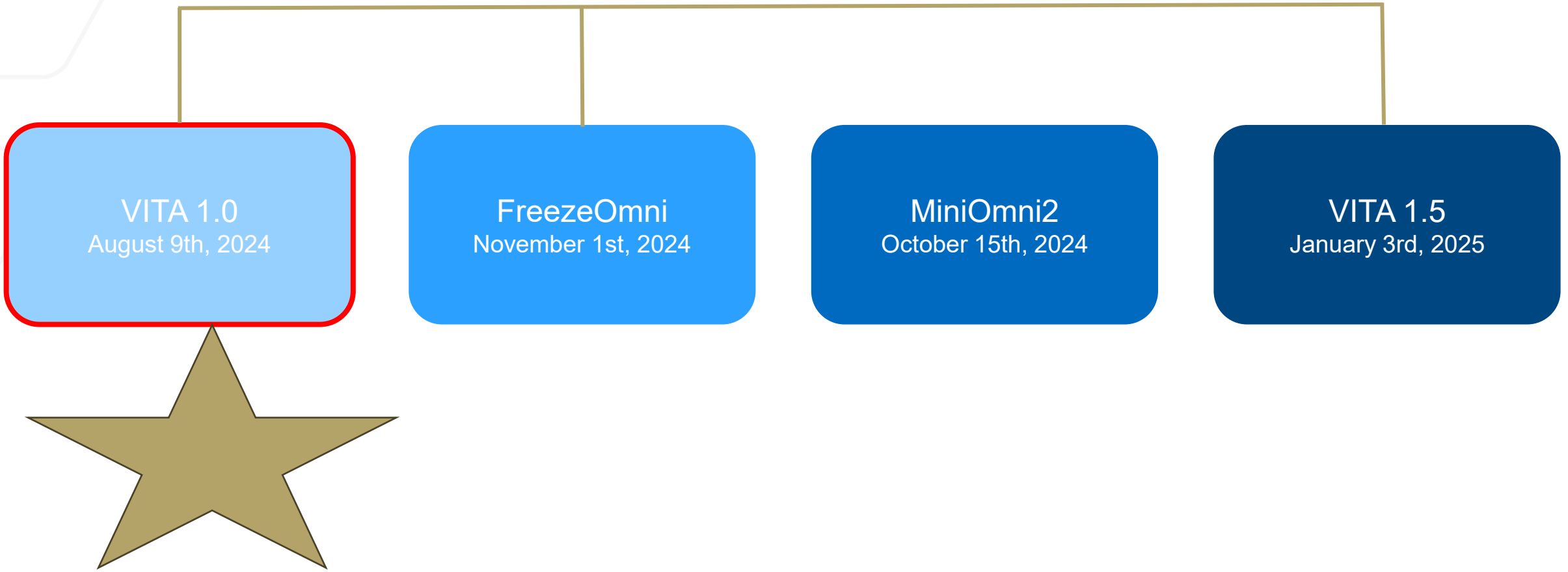
Chengyin Xu
MS Computer Science
cxu371@gatech.edu

Efficient ML, Graph ML

Problem Statement

- **MLLMs excel at vision–language tasks**, but struggle to extend these capabilities to **speech**, especially for open-sourced models.
- **Speech integration is nontrivial** since it encodes **temporal dynamics**, while vision encodes **spatial structure**, causing optimization conflicts across modalities.
- **Conventional speech pipelines** rely on modular Automatic Speech Recognition (ASR) + Text to Speech (TTS) systems, which introduce **latency**, **loss of coherence**, and **fragmented learning**.
- **Core challenge:** How to achieve *end-to-end multimodal understanding* and *real-time speech interaction* without degrading visual reasoning?
- **VITA-1.5** tackles this via a **progressive three-stage training strategy** that incrementally aligns **vision, language, and speech**, while preserving performance in each modality.

History



Fu et al, VITA: Towards Open-Source Interactive Omni Multimodal LLM https://arxiv.org/html/2408.05211v1
Wang et al, Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM https://arxiv.org/html/2411.00774v1
Xie et al, Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities https://arxiv.org/abs/2410.11190
Fu et al, VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction https://arxiv.org/abs/2501.01957

VITA: Towards Open-Source Interactive Omni Multimodal LLM

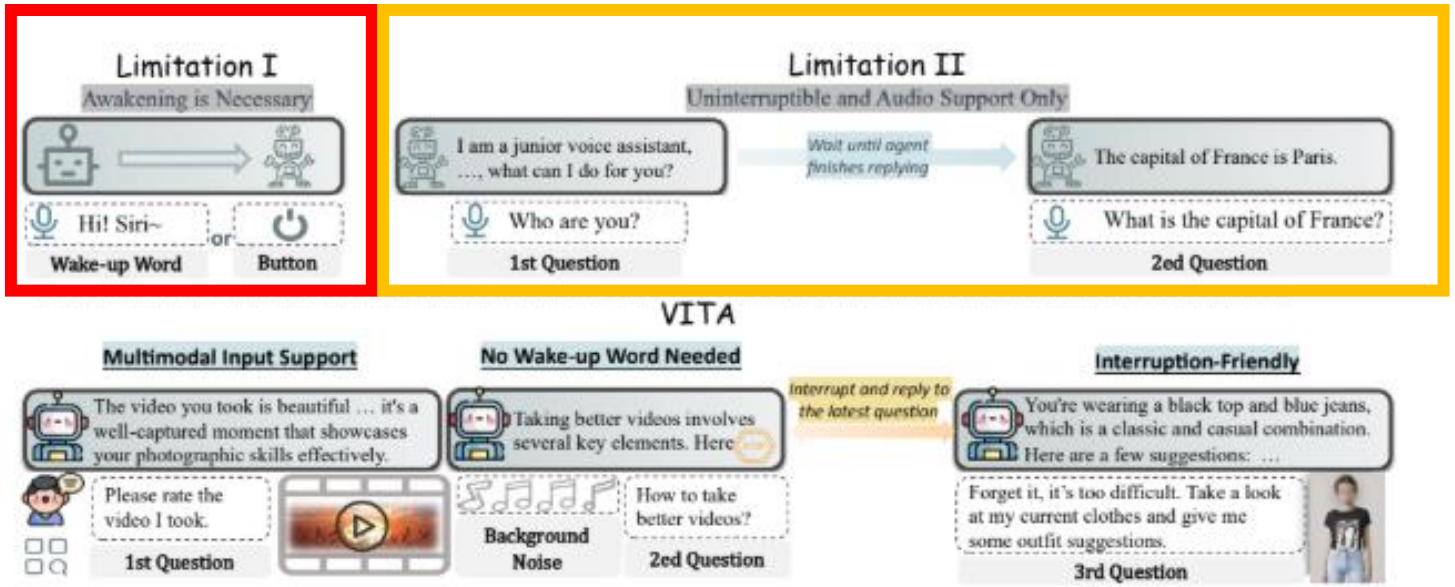
Georgia Tech

Problem Statement

VITA: simultaneous processing of **V**ideo, **I**mage, **T**ext, and **A**udio modalities

Previous Methods:

1. Awakening is necessary
2. Uninterruptible
3. Audio Support Only
4. Lack of Open-Source Models



Georgia Tech

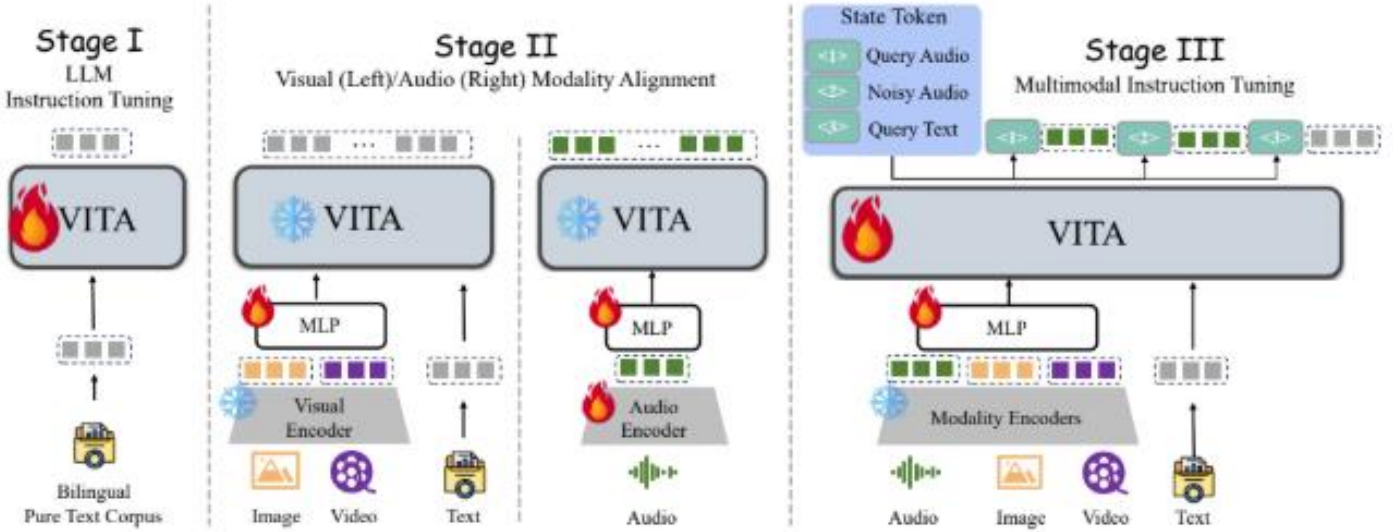
Training Scheme

Base Model: Mixtral 8x7B

1. Stage I:
   Bilingual Instruction Tuning of LLM

2. Stage II:
   Multimodal Alignment

3. Stage III:
   Multimodal Instruction Tuning

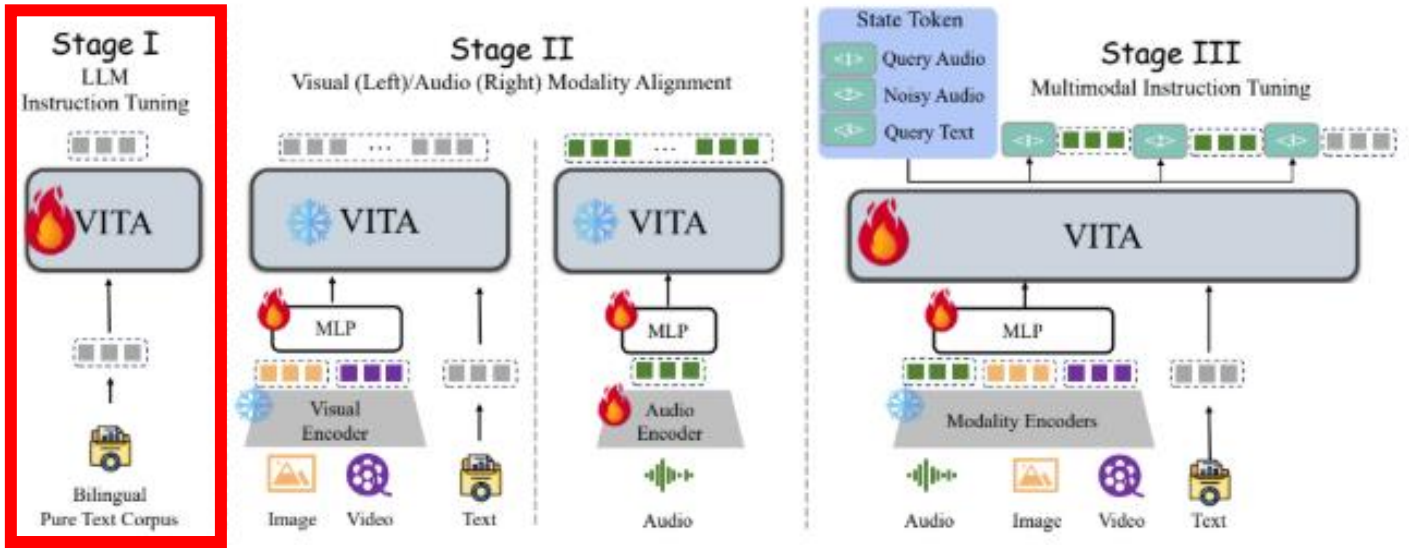


Georgia Tech

Training Scheme

Base Model: Mixtral 8x7B

1. Stage I:
   Bilingual Instruction Tuning of LLM



Issue:
Mixtral shows limited proficiency in understanding Chinese.

Goal:
Enable bilingual (Chinese–English) skill.

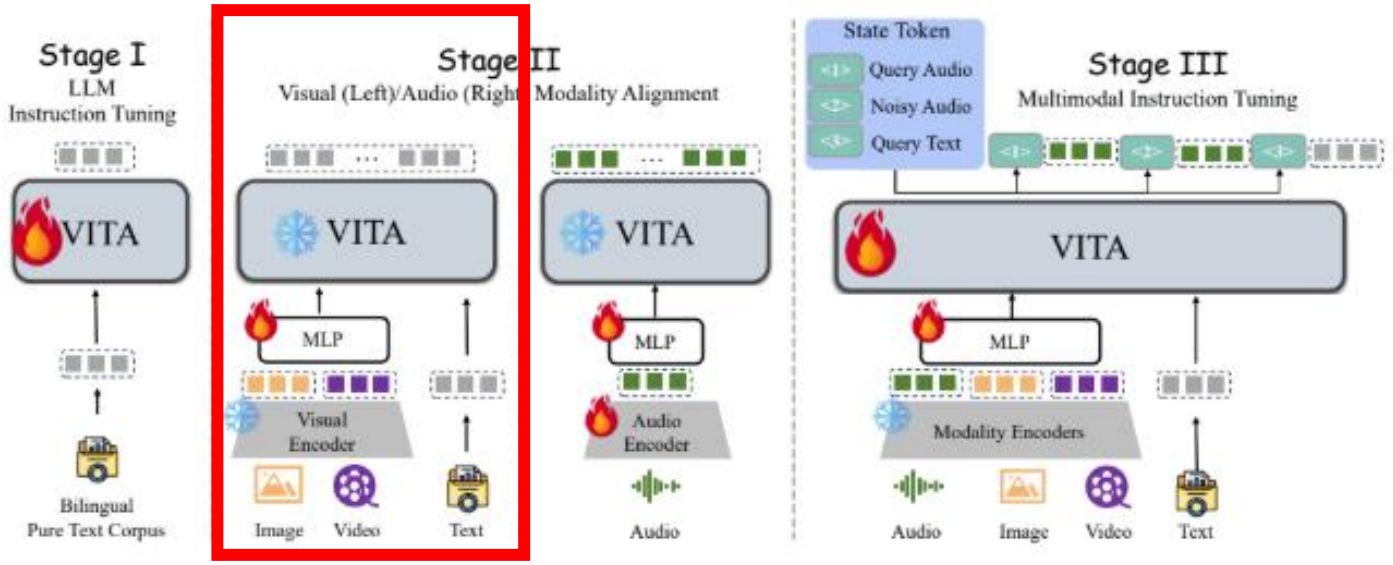
Step 1 — Vocabulary expansion:
Extend the base model's vocab from **32,000** → **51,747** to include Chinese tokens.
Benefit of expansion:
Fewer tokens per Chinese text segment → **higher inference efficiency**

Step 2 — Training data:
Perform **pure-text instruction tuning** using **5M synthetic bilingual (Zh–En) pairs**.
Outcome:
Improved Chinese understanding while preserving English performance.

Training Scheme

Base Model: Mixtral 8x7B

2. Stage II:
   Multimodal Alignment



Visual Alignment:
Backbone: InternViT-300M (448px input resolution).

Tokenization (images): 448×448 image
→ 256 visual tokens via a 2-layer MLP connector
→ High-res images: Use dynamic patching

Videos treated as multi-image:
1) < 4 s: uniformly sample 4 frames total
2) 4–16 s: sample 1 frame per second
3) > 16 s: uniformly sample 16 frames total

No dynamic patching on video frames (prevents token explosion and keeps latency manageable).

Token budget intuition: per frame ≈ 256 tokens → total video tokens ≈ 256 × (sampled frames).

Exactly same in VITA 1.5!

Georgia Tech

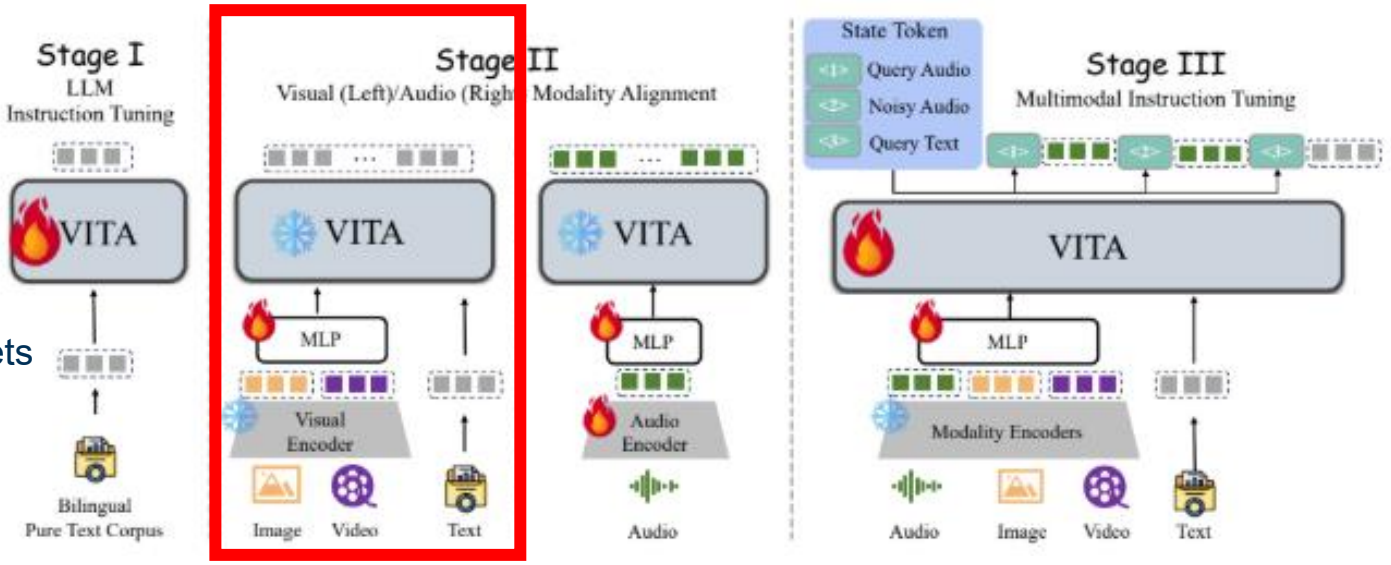
Training Scheme

Base Model: Mixtral 8x7B

2. Stage II:
   Multimodal Alignment

Visual Alignment:

Train **Visual Connector** on Image Description and VQ datasets
Visual Encoder, VITA (LLM) frozen, No Audio
Data Concatenation to 6K tokens for computational efficiency

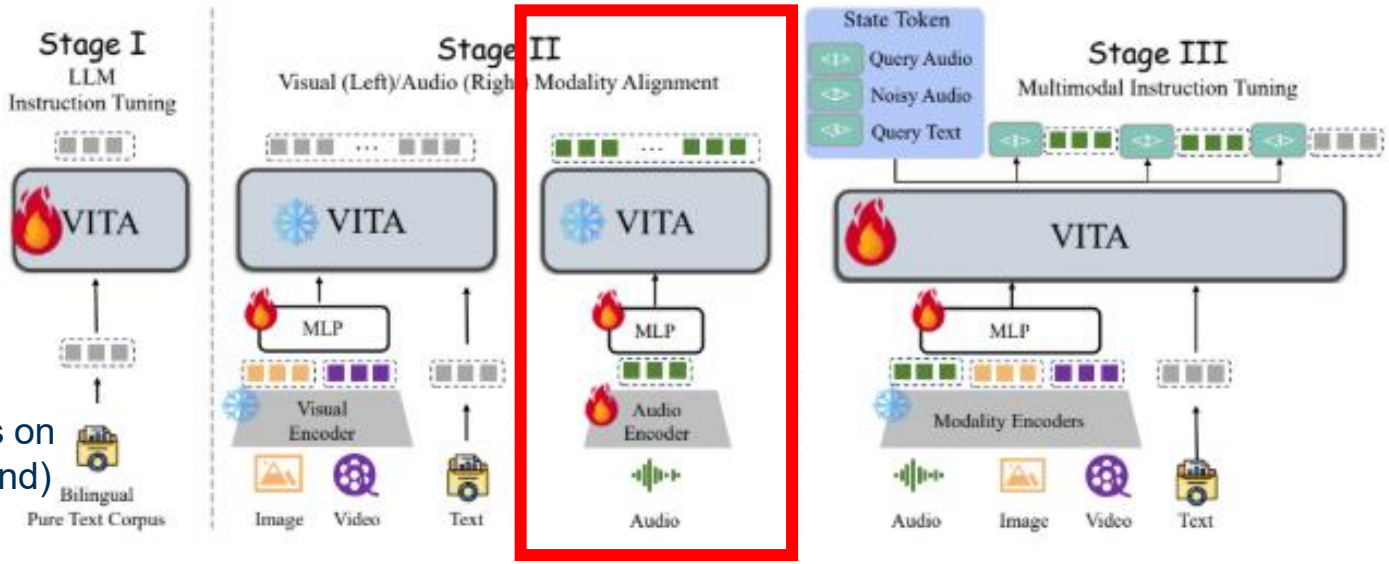


Georgia Tech

Training Scheme – Until Here

Base Model: Mixtral 8x7B

2. Stage II:
   Multimodal Alignment



Audio Modality:

Front-end features:

Log-mel filterbank (mel-scale)

Breaks down the audio signal into individual frequency bands on the mel frequency scale (nonlinear human perception for sound)

Encoder stack (~341M Params):

4× CNN downsampling (reduces time resolution)

24-layer Transformer

Connector:

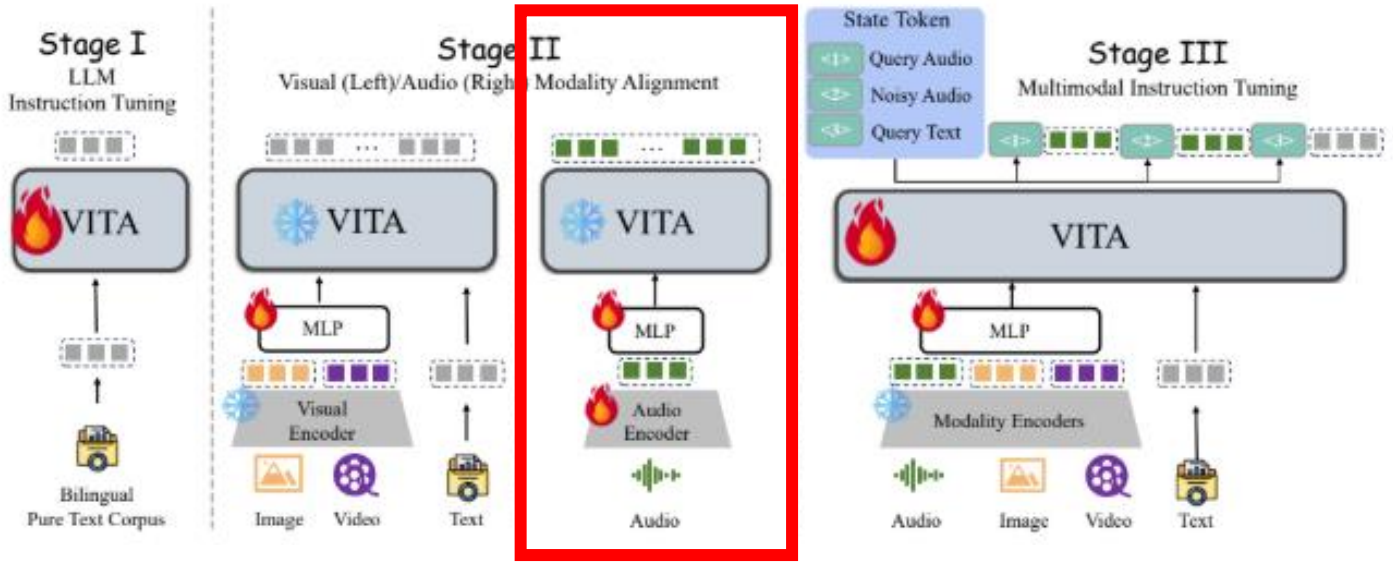
2-layer MLP: maps audio features into the LLM token space.

Exactly same in VITA 1.5!

Training Scheme

Base Model: Mixtral 8x7B

2. Stage II:
   Multimodal Alignment



Audio Modality:
Token rate: ~12.5 Hz → 25 tokens per 2 seconds of audio.

Audio alignment tasks
Trained components: encoder + connector

ASR objective:
WenetSpeech (~10k hours; mainly Chinese, multi-domain).
GigaSpeech (~10k hours; mainly English, high-quality).

Audio captioning objective:
AudioSet-SL subset of WavCaps (~400k clips with captions).

Purpose of alignment: teach the model to map audio → text space (transcribe/describe) so the LLM can reliably consume audio-conditioned tokens during downstream multimodal instruction tuning.

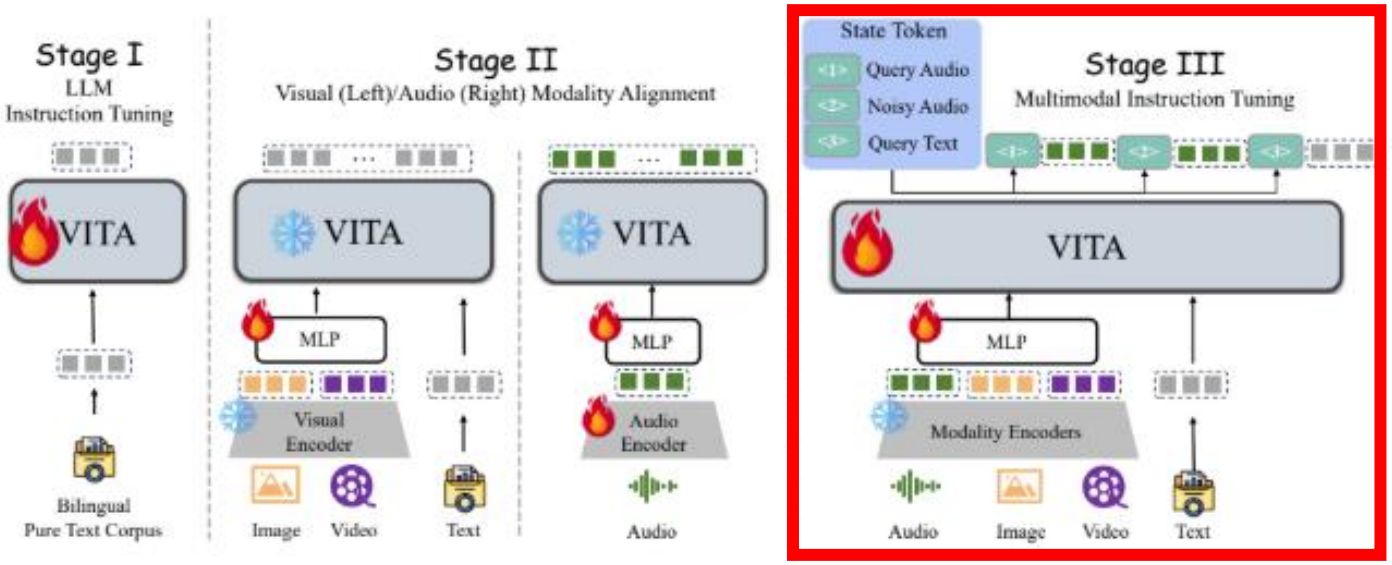
Training Scheme

Base Model: Mixtral 8x7B

3. Stage III:
   Multimodal Instruction Tuning

Training Data

Text↔Speech mix:
Randomly replace ~50% of text questions with TTS-spoken versions (e.g., GPT-SoVITS) to create audio queries.



System prompts:
Distinct prompts per data type

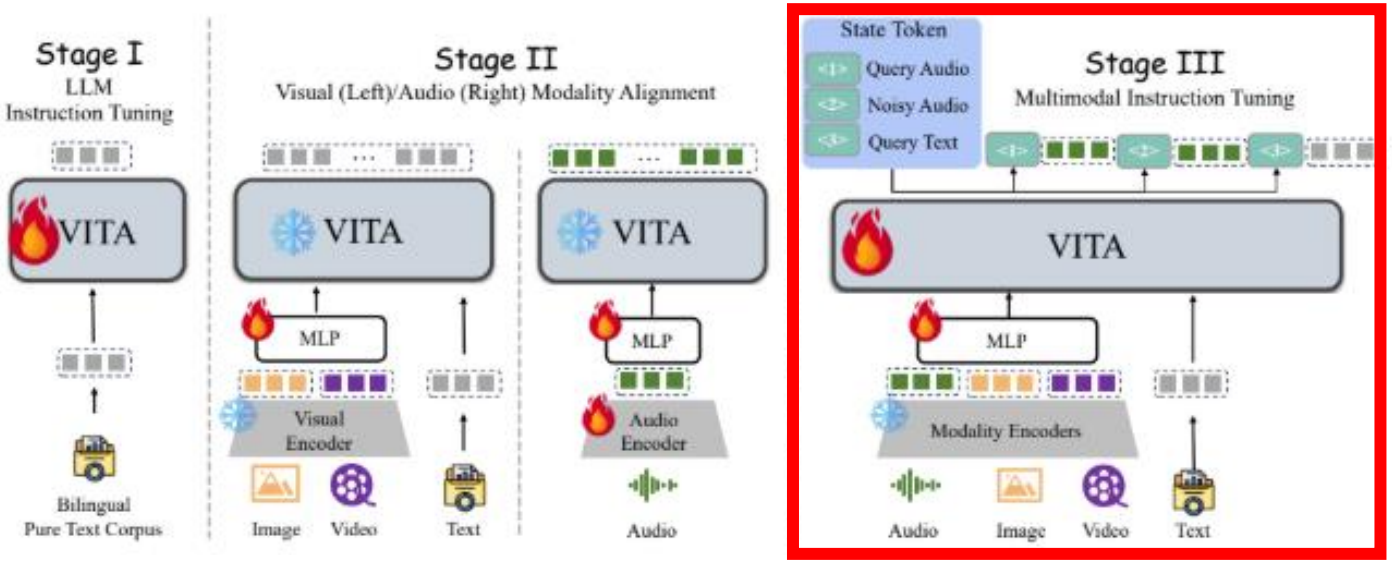
Noisy audio construction:
Synthesize **non-query** clips by TTS'ing negative sentences to teach **ignore/silent** behavior.

Georgia Tech

Training Scheme

Base Model: Mixtral 8x7B

3. Stage III:
   Multimodal Instruction Tuning



Training Data

State tokens (input control):

- <1> **Query Audio** → answer
- <2> **Noisy/Non-query Audio**

→ trained with "noisy" text targets; at runtime treated as **EOS/no-reply**;

- <3> **Query Text** → answer.

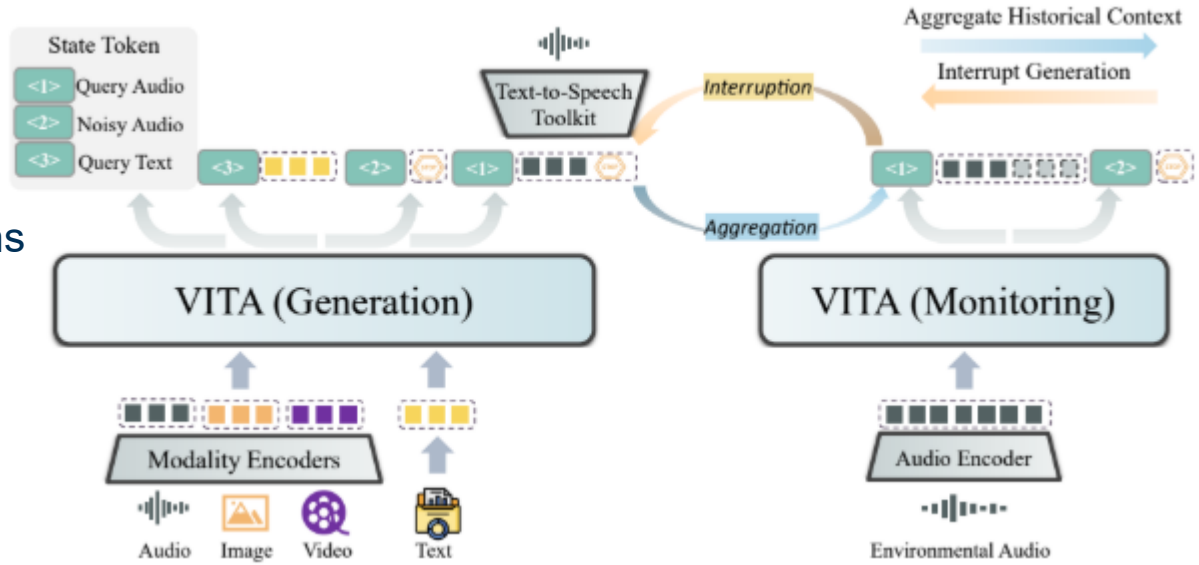
Implementation

Duplex Pipeline

Non-awakening Interaction:
The model can be activated and respond to user audio questions in the environment without the need for a wake-up word or button.

1) Real-time Tracking of Environmental Sounds
-VAD: Voice Activity Detection (SileroVAD)

2) Filtering out noisy audio.
-The model should only respond to effective human query audio.
-State Token <2>: If the input is of a non-query type, the model directly terminates the inference

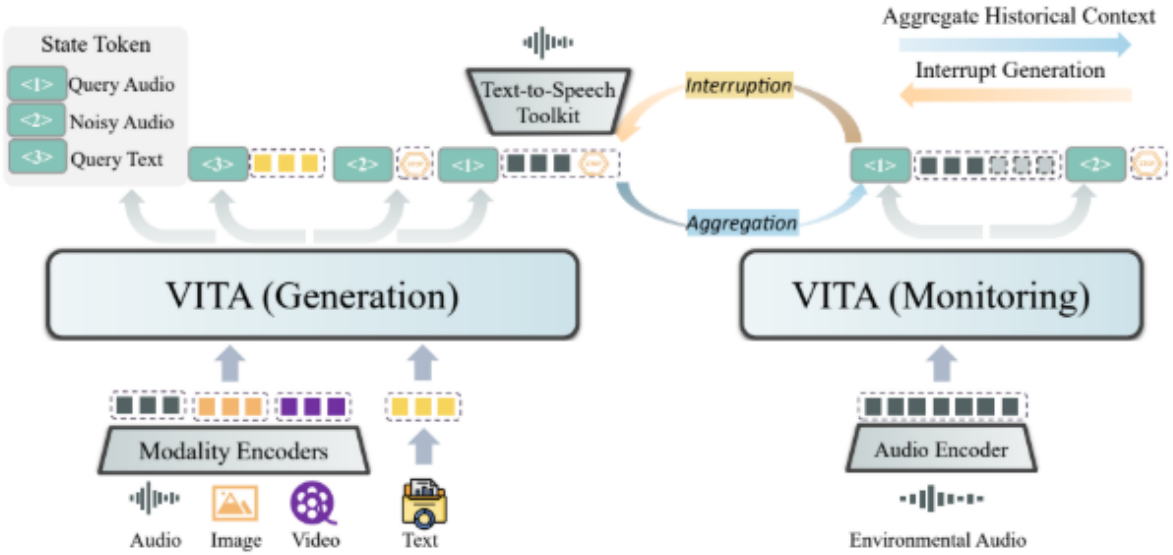


Implementation

Duplex Pipeline

Audio Interrupt Interaction:

Enables users to interrupt the model's generation at any time with new questions.



- Real-time Tracking and Filtering of External Queries:
While generating responses, the system must simultaneously track and filter external queries in real time.

- Answering New Questions:
When a new question emerges, the system must cease its current generation, consolidate the historical context, and respond to the present query.

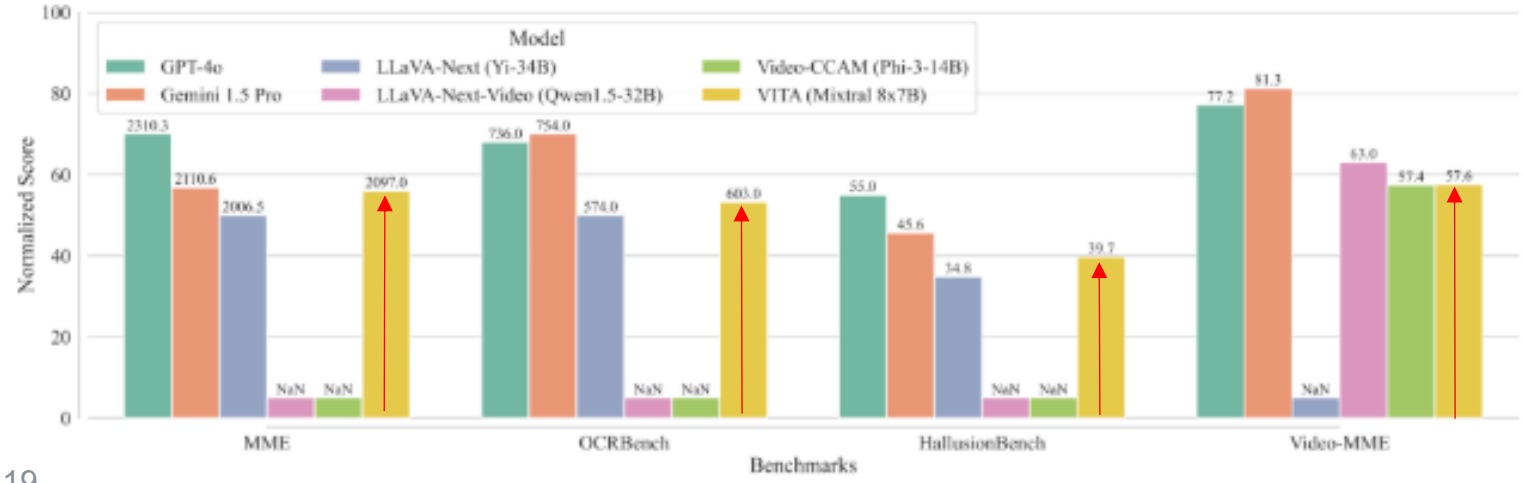
Duplex Pipeline:

1. Two VITA models are deployed concurrently.
2. Under a typical condition, the Generation model answers user queries. Simultaneously, the Monitoring model detects environmental sounds during the generation process.
3. The Monitoring model disregards non-query user sounds, i.e., noisy audio, but ceases the Generation model's progress when it identifies query audio.
4. When ceased, the Monitoring model subsequently consolidates the historical context and responds to the latest user query.

Results

Method	C-EVAL	AGIEVAL	MMLU	GSM8K
	CN	CN & ENG	ENG	ENG
Mixtral-8x7B Instruct	53.30	41.72	70.35	63.99
Mixtral-8x7B Ours	56.68	46.17	70.98	75.66

Method	Wenetspeech (CN)		Librispeech (ENG)			
	Test_Net	Test_Meeting	Dev_clean	Dev_other	est_clean	est_other
VITA	12.15	16.53	7.57	16.57	8.14	18.41



WER: Word Error Rate
What it measures: how many **word-level** edits (substitutions S, deletions D, insertions I) your ASR needs to turn its hypothesis into the reference.

WER = (S+D+I) / N_words

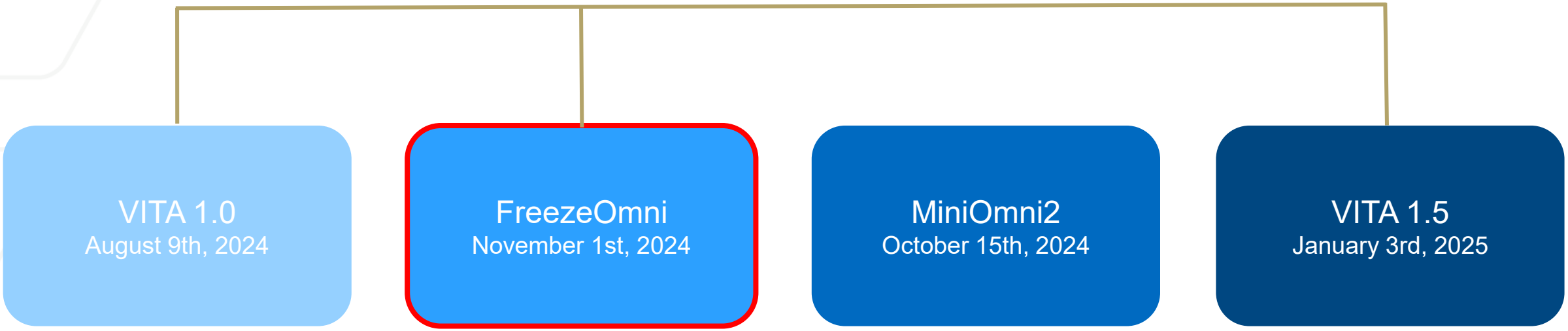
Example (English):
Ref: "the cat sat on the mat" (6 words)
Hyp: "the cat sat on mat"
Edits: 1 deletion ("the" before "mat") → S=0,D=1,I=0
WER = 1/6 ≈ 16.7%.

CER: Character Error Rate
What it measures: same idea, but at the **character** level (good for languages without spaces, like Chinese).

CER = (S+D+I) / N_chars

Example (English):
Ref: "kitten" (6 chars)
Hyp: "sitting"
Optimal edits (Levenshtein): k→s (S=1), e→i (S=1), add "g" (I=1) → total 3 edits.
CER = 3/6 = 50%.

History



Fu et al, VITA: Towards Open-Source Interactive Omni Multimodal LLM https://arxiv.org/html/2408.05211v1
Wang et al, Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM https://arxiv.org/html/2411.00774v1
Xie et al, Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities https://arxiv.org/abs/2410.11190
Fu et al, VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction https://arxiv.org/abs/2501.01957

Problem Statement

For interaction with LLM in Speech Modality,

Traditional method :

"Use a cascaded approach of ASR + LLM + TTS"

Limitation :

- High engineering complexity
- High interaction latency.

New method:

"By the parameters of the LLM are more or less fine-tuned, Aligning the LLM with the speech modality"

Limitation :

- The forgetting problem to the LLM, resulting in a negative impact on its intelligence
- an obvious gap in performance between spoken question-answering and text-modality question-answering

Freeze-Omni suggests:

"Achieving speech modality alignment while the LLM is frozen throughout the training process, and obtaining low latency speech dialogue capabilities while keeping the intelligence of the backbone LLM."

Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM

Xiong Wang[1,*], Yangze Li[2], Chaoyou Fu[3], Lei Xie[2], Ke Li[1], Xing Sun[1], Long Ma[1,†]

[1]Tencent Youtu Lab
[2]Audio, Speech and Language Processing Group (ASLP@NPU)
[3]Nanjing University
[*] Main Contribution [†] Corresponding Author

https://freeze-omni.github.io/

Abstract

The rapid development of large language models has brought many new smart applications, especially the excellent multimodal human-computer interaction in GPT-4o has brought impressive experience to users. In this background, researchers have proposed many multimodal LLMs that can achieve speech-to-speech dialogue recently. In this paper, we propose a speech-text multimodal LLM architecture called **Freeze-Omni**. Our main contribution is the speech input and output modalities can connected to the LLM while keeping the LLM frozen throughout the training process. We designed 3-stage training strategies both for the modeling of speech input and output, enabling Freeze-Omni to obtain speech-to-speech dialogue ability using text-speech paired data (such as ASR and TTS data) and **only 60,000 multi-round text Q&A data on 8 GPUs**. Moreover, we can effectively ensure that the intelligence of the Freeze-Omni in the speech modality is at the **same level** compared with that in the text modality of its backbone LLM, while the end-to-end latency of the spoken response achieves a low level. In addition, we also designed a method to achieve duplex dialogue ability through multi-task training, making Freeze-Omni have a more natural style of dialogue ability between the users. Freeze-Omni mainly provides a possibility for researchers to conduct multimodal LLM under the condition of a frozen LLM, avoiding various impacts caused by the catastrophic forgetting of LLM caused by fewer data and training resources.

1 Introduction

In recent years, the development of large language models has been extremely rapid. A series of large language models represented by the GPT series [10, 1] of OpenAI has demonstrated extraordinary capabilities. As speech interaction is one of the most natural forms of human-computer interaction, combining speech input and output with an LLM can bring an extraordinary experience to users. The traditional method is to use a cascaded approach of ASR + LLM + TTS to achieve the interaction with LLM in speech modality. However, this approach often leads to a relatively high engineering complexity and a considerable interaction latency. Nevertheless, GPT-4o [18] has changed this situation, it provides an end-to-end speech interaction mode which has significantly improved the user experience, triggering a research boom among researchers regarding multimodal LLMs for speech-to-speech interaction.

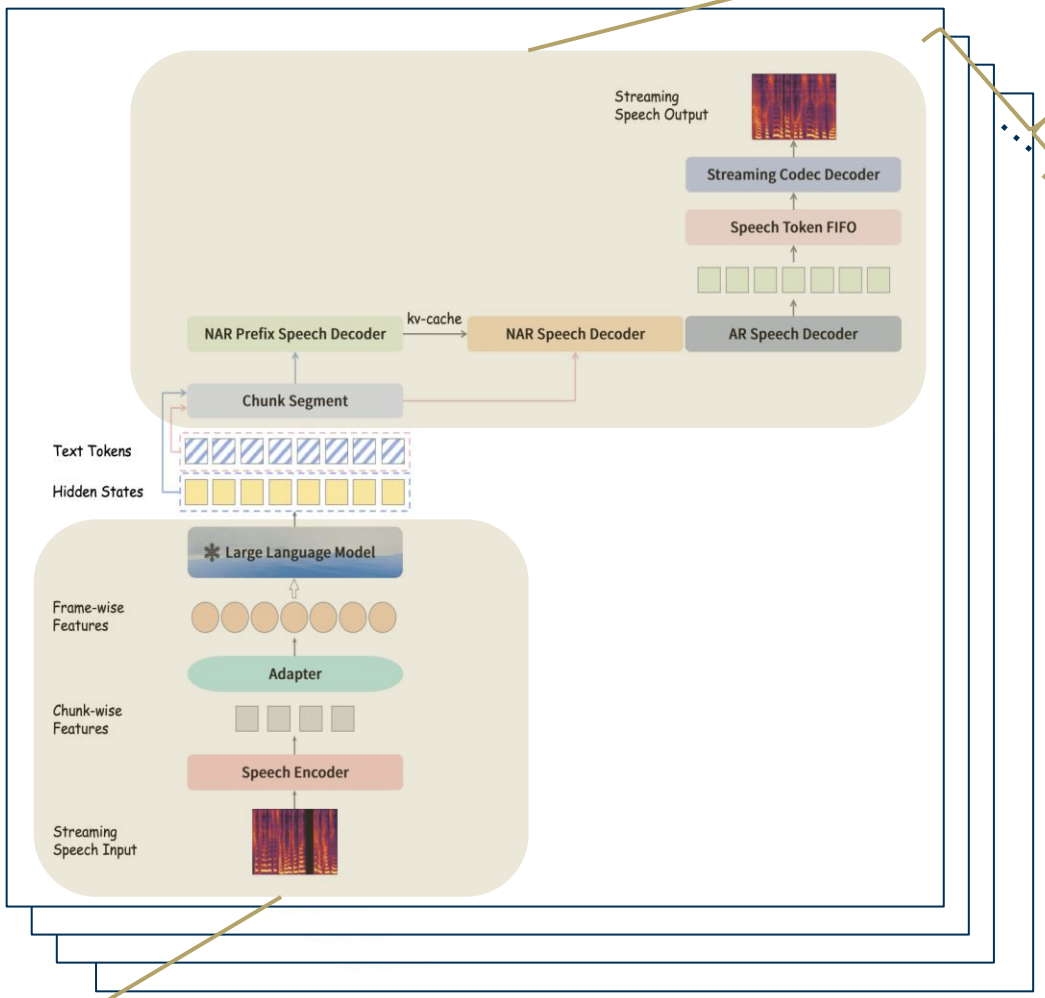
In the field of general LLMs, many public models such as Llama 3.2 [8], Qwen2.5 [21], Mixtral [14], etc. have provided very good opportunities for researchers to develop downstream tasks on them. Therefore, in the research field of multimodal LLMs for speech-to-speech, works such as Mini-Omni2 [24], LLaMA-Omni [9], and Moshi [7] have provided excellent references for researchers.

Email: wangxiongts@gmail.com, malonema@tencent.com

Georgia Tech

Overview

Modeling of Speech Output



Design for Duplex Dialogue

Modeling of Speech Input

Georgia Tech

Overview

Modeling of Speech Input

1) The alignment between the speech input to text output,

Modeling of Speech Output

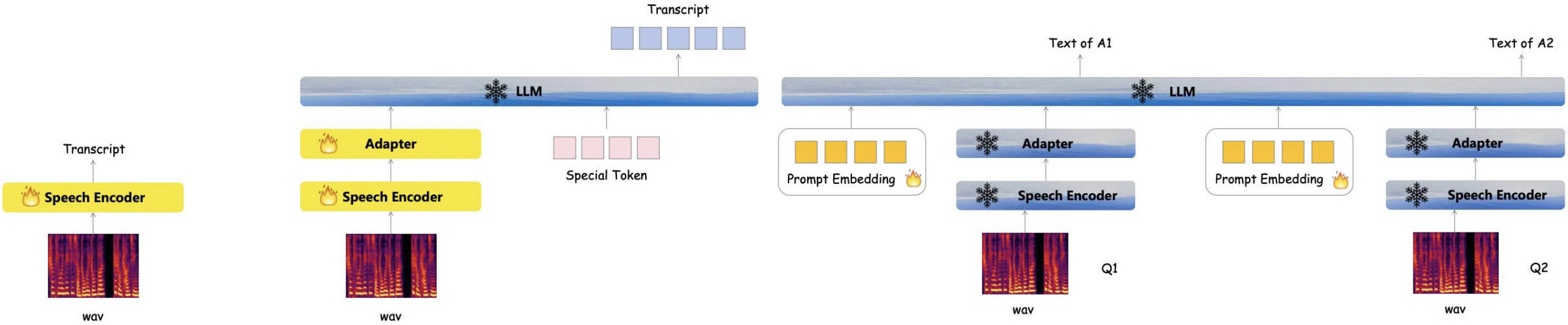
2) The alignment between the text input to speech output

Design for Duplex Dialogue

3) By connecting these two components with the LLM

The Ability of Speech Input to Speech Output

Modeling of Speech Input



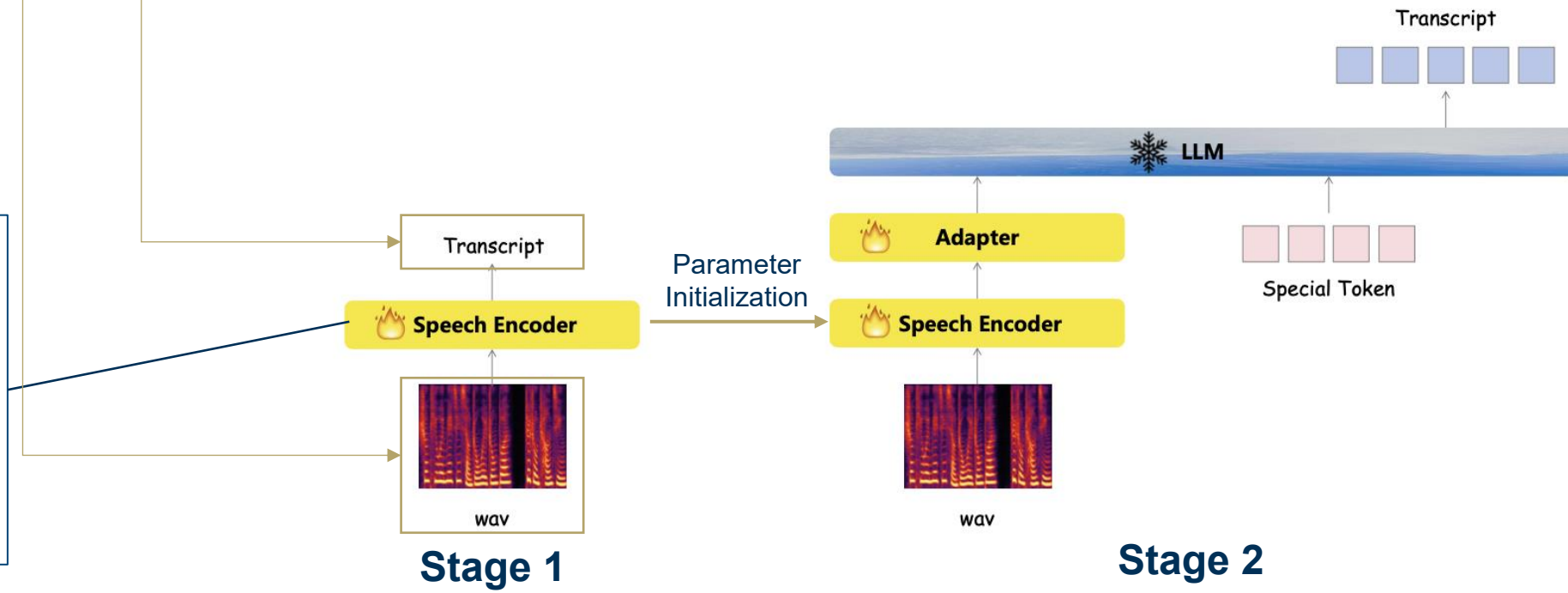
Georgia Tech

Modeling of Speech Input

Dataset : 110,000h internal speech-text paired ASR data including both Chinese and English

Utilizes a chunk-wise streaming speech encoder to transform the input speech features into a high-dimensional representation.

- A multi-layer convolution with 4-times down sampling and 24 layers of transformers with a hidden size of 1024.
- # of Parameter : 350M



Training :

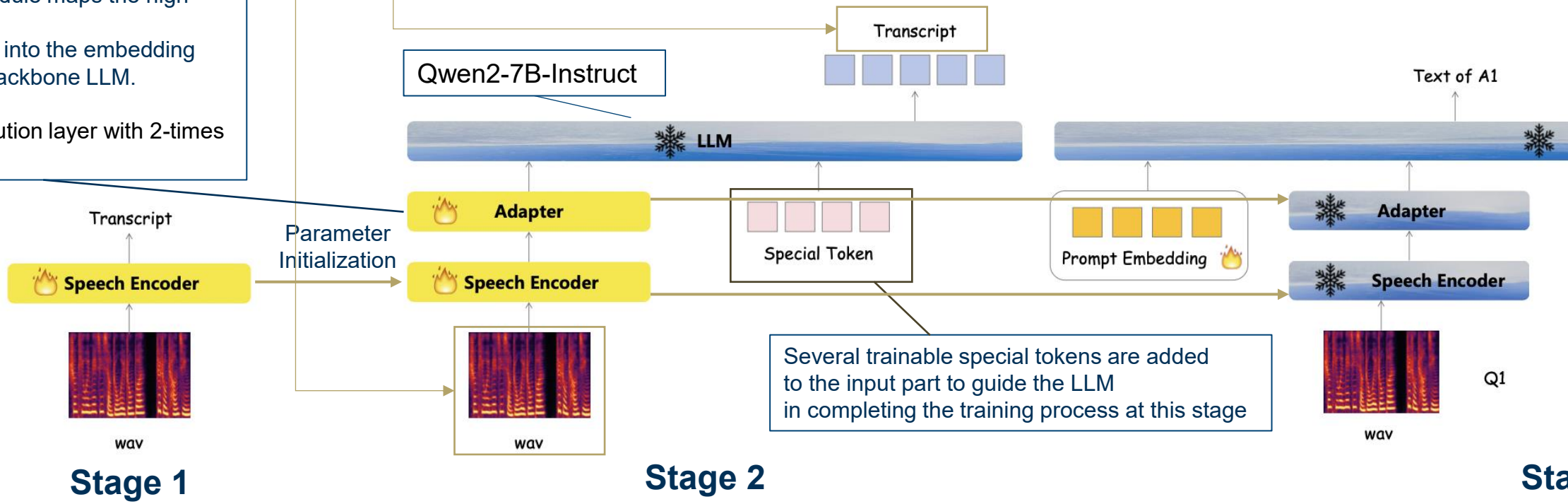
- Loss Function: CTC
- Optimizer : Adamw [16] optimizer with a warm-up learning rate scheduler
- Learning Rate:
  - Stage 1: 2e-4

Modeling of Speech Input

Dataset : 110,000h internal speech-text paired ASR data including both Chinese and English

an adapter module maps the high-dimensional
representation into the embedding space of the backbone LLM.

A multi-convolution layer with 2-times downsampling



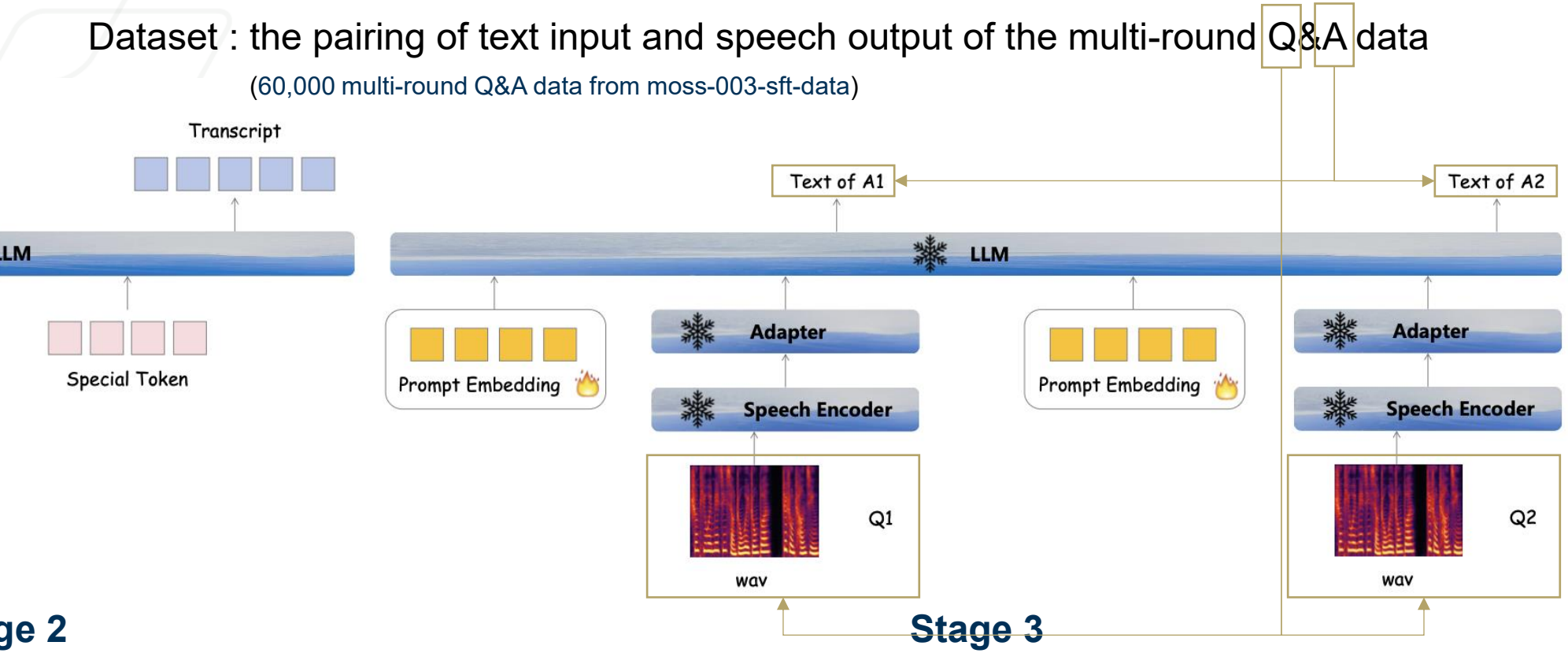
Training :

- Loss Function: CTC
- Optimizer : Adamw [16] optimizer with a warm-up learning rate scheduler
- Learning Rate:
  - Stage 1: 2e-4
  - Stage 2: 1e-4

Modeling of Speech Input

Dataset : the pairing of text input and speech output of the multi-round Q&A data

(60,000 multi-round Q&A data from moss-003-sft-data)



ge 2

Training :

- Loss Function: CTC
- Optimizer : Adamw [16] optimizer with a warm-up learning rate scheduler
- Learning Rate:
  - Stage 1: 2e-4
  - Stage 2: 1e-4
  - Stage 3: 6e-4

Results on Speech Input

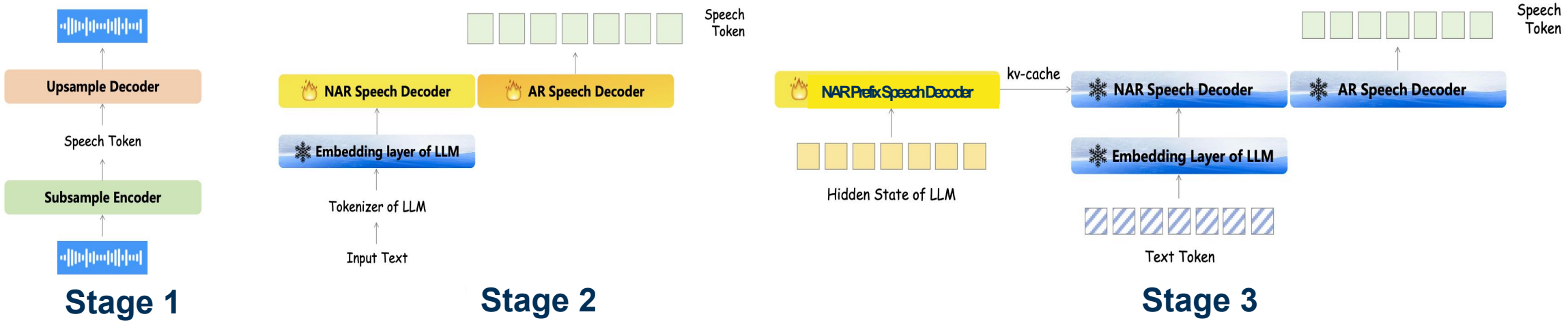
Table 1: The ASR performance of the model corresponding to stage 2 in the modeling of speech input, where {aishell-1 [4],test_net [26], test_meeting [26]} are Mandarin evaluation sets, measured in CER (%), while {dev-clean,dev-other,test-clean,test-other} [19] are English evaluation sets, measured in WER (%).

Model	aishell-1	test_net	test_meeting	dev-clean	dev-other	test-clean	test-other
Wav2vec2-base [2]	-	-	-	6.0	13.4	-	-
Mini-Omni2 [24]	-	-	-	4.8	9.8	4.7	9.4
Freeze-Omni							
+ $chunk = \infty$	2.15	8.57	10.09	3.29	7.4	3.24	7.68
+ $chunk = 4$	2.79	12.6	14.2	4.16	10.21	4.05	10.48
+ w/o dynamic	2.48	11.8	13.46	4.03	9.45	3.82	9.79

Dynamic Chunk Training enables the model **to handle both streaming and offline conditions** by training with variable chunk sizes.

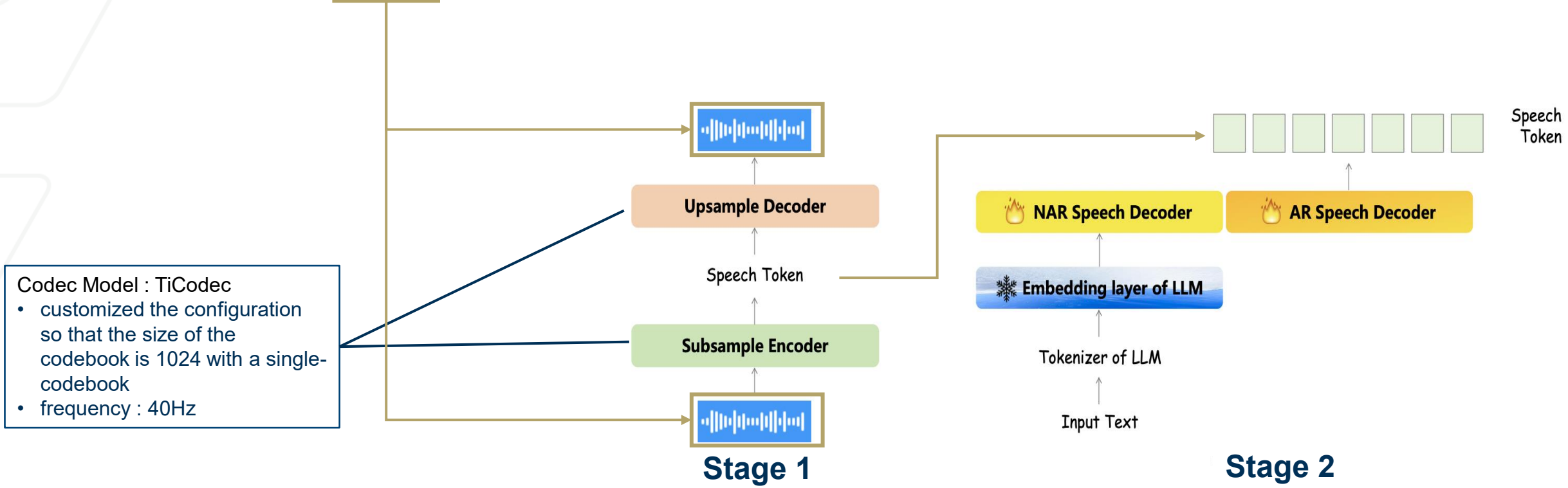
Although a fixed chunk (e.g., 4) yields slightly lower error rates without dynamics, the **dynamic approach offers far better generalization and flexibility**.

Modeling of Speech Output



Modeling of Speech Output

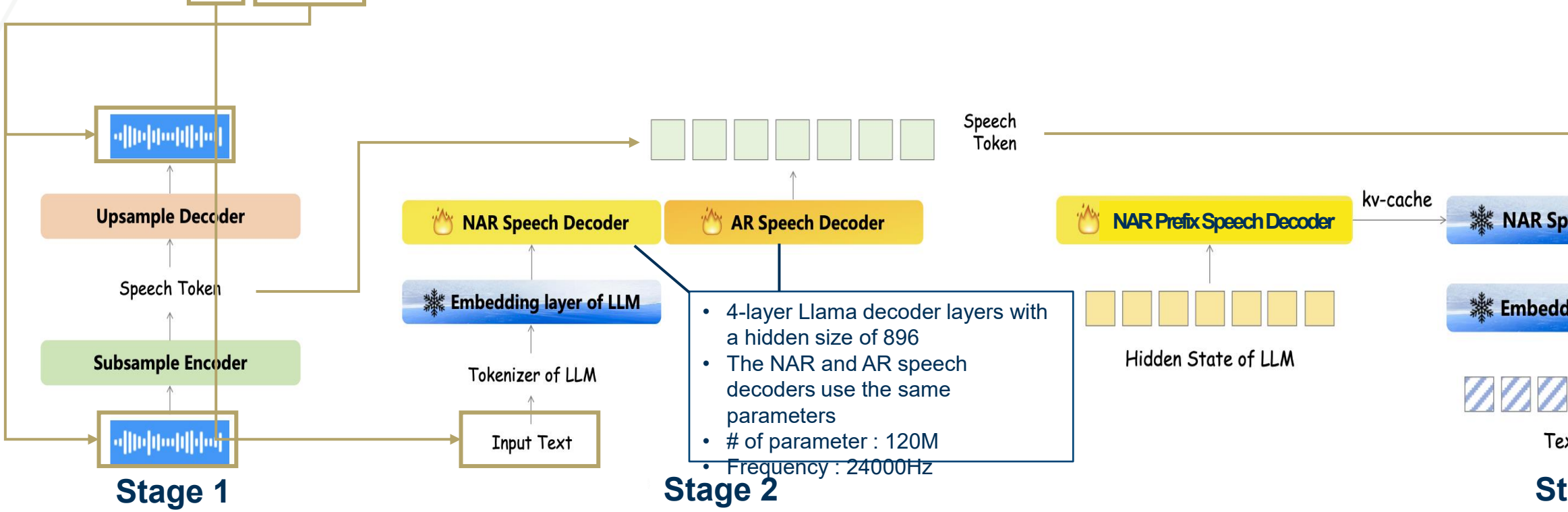
Dataset : 3,000h of text-speech paired data generated by a zero-shot TTS system



Georgia Tech

Modeling of Speech Output

Dataset : 3,000h of text-speech paired data generated by a zero-shot TTS system



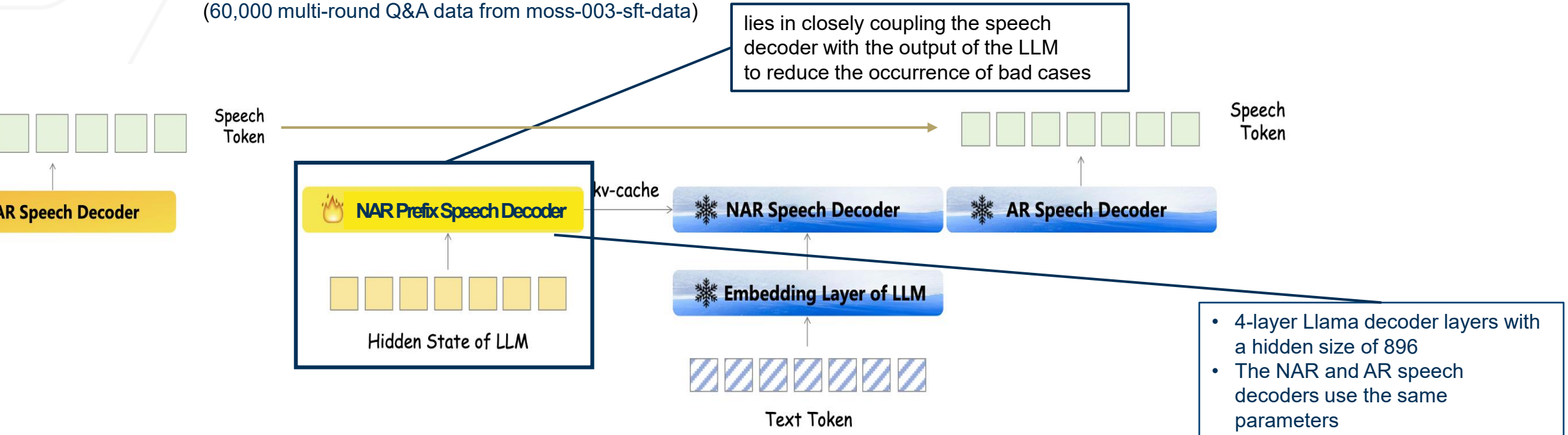
Training :

- Loss Function: CTC
- Optimizer : Adamw [16] optimizer with a warm-up learning rate scheduler
- Learning Rate:
  - Stage 1: Same as TiCodec
  - Stage 2: 5e-5

Modeling of Speech Output

Dataset : the pairing of text input and speech output of the multi-round Q&A data

(60,000 multi-round Q&A data from moss-003-sft-data)



e 2

Stage 3

Training :

- Loss Function: CTC
- Optimizer : Adamw [16] optimizer with a warm-up learning rate scheduler
- Learning Rate:
  - Stage 1: Same as TiCodec
  - Stage 2: 5e-5
  - Stage 3: 5e-5

Results on speech output

Experimental Setup

- 1,000 utterances were randomly sampled, using **text tokens and hidden states** from the LLM as inputs to the speech decoder.
- The generated speech was evaluated by **ASR accuracy (CER%)** using *paraformer-zh*, under different **top-k decoding** settings.
- Two models were compared: **Speech Decoder w/o Prefix** (stage 2) and **Speech Decoder with Prefix NAR** (stage 3).

Table 2: The CER(%) of the speech decoder on 1,000 evaluation utterances under different *top-k*.

Method	*top-k* 1	2	3	4	5
Speech Decoder w/o Prefix	5.27	4.64	4.76	4.66	5.03
Speech Decoder	3.9	3.65	3.53	3.62	3.71

Prefix NAR Decoder **improves alignment** with the LLM and **reduces CER**, showing better robustness across all **top-k** values.

Design for duplex dialogue

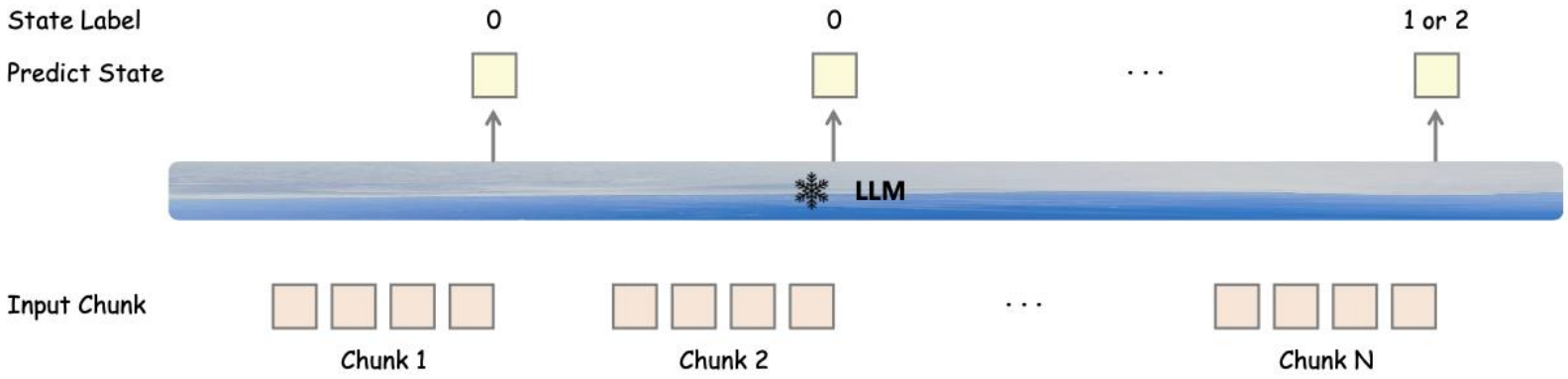
use multi-task for chunk-level state prediction

State 0 : the current LLM can continue to receive speech

State 1 : the LLM can interrupt the user and perform the generate stage

State 2 : there is no need to interrupt the user

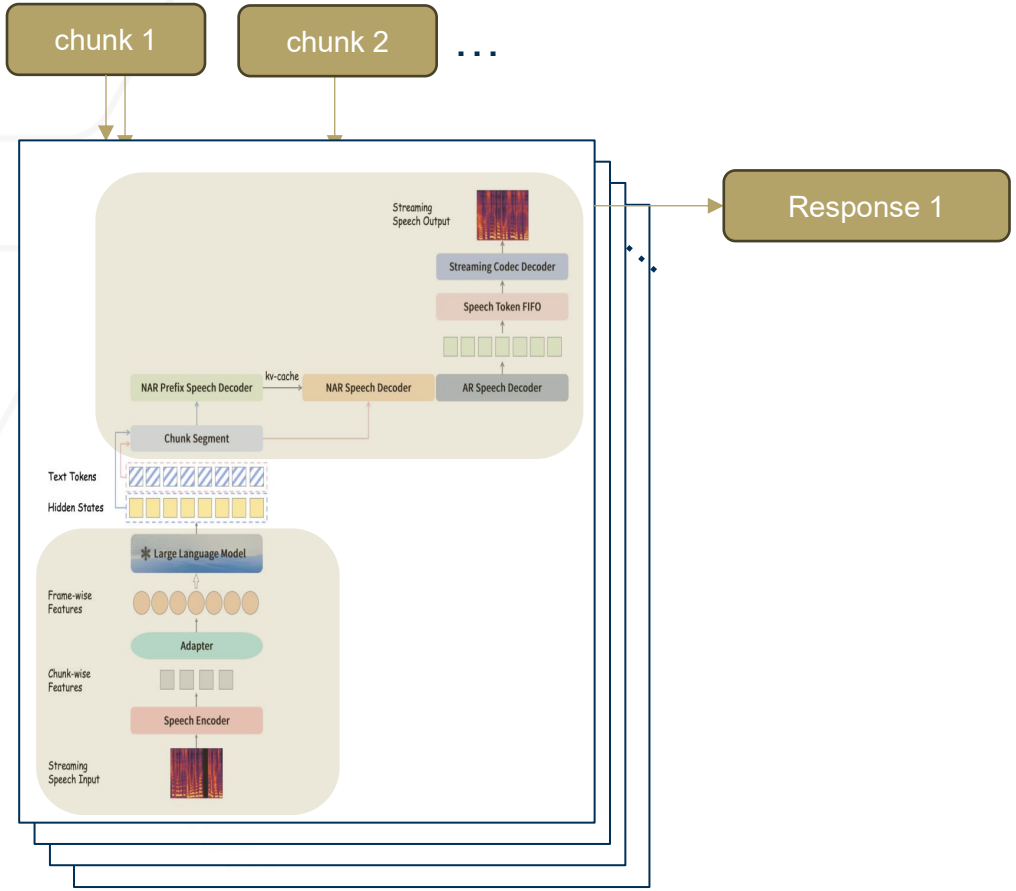
Stop sending speech streams to Freeze-Omni and reset the VAD module.



Using a multi-task method to optimize the cross-entropy loss of both the state classification layer and the LLM.

Design for duplex dialogue

"model as a server" strategy for the speech-to-speech dialogue system



- Started several models simultaneously
- A user's VAD was triggered, the speech would be sent to the server in the form of chunks
- Server would be responsible for scheduling which idle model should respond to the current chunk
- Separated all the kv-cache and CNN cache of the speech encoder and LLM

"Any model in the server could respond to any chunk of any user"

Results on spoken question answering

Table 3: The accuracy (%) of different models in question answering on three sets. The models in the first four rows all use speech as input, while the models in the last two rows use text as input. The backbone LLM of Freeze-Omni is Qwen2-7B-Instruct, and the backbone LLM of Moshi is Helium. Both Freeze-Omni and Qwen2-7B-Instruct use greedy search in the generate stage with zero-shot, and the accuracy is calculated using the output text. Except for Freeze-Omni and Qwen2-7B-Instruct, previous evaluation results are derived from corresponding references.

Model	Modality	Web Q.	LlaMA Q.	Audio Trivia QA
SpeechGPT(7B) [27]	Audio&Text	6.5	21.6	14.8
Spectron(1B) [17]	Audio&Text	6.1	22.9	-
Moshi(7B) [7]	Audio&Text	26.6	62.3	22.8
Freeze-Omni(7B)	Audio&Text	44.73	72	53.88
Helium [7]	Text Only	32.3	75	56.4
Qwen2-7B-Instruct	Text Only	45.13	77.67	63.93

- **Freeze-Omni** achieves **much higher accuracy** than other speech-based models.
- Its performance is **close to text-only LLMs (Qwen2-7B-Instruct)**, showing strong alignment between **speech and text understanding**.
- Demonstrates that **Freeze-Omni preserves reasoning ability even from spoken input**.

Analysis on end-to-end latency

- **Statistical latency:** Time from LLM interruption to the first generated speech PCM chunk (automatically measurable).
- **Non-statistical latency:** Real-time delay from the end of user speech to LLM's response start (manually measured).

Table 4: Detailed information of statistical latency. Among them, 50% represents the median, and 90% represents the percentile at 90. The unit of the results in the table is (ms). All results are completed using pytorch with bfloat16 inference on a single NVIDIA A100 GPU.

Latency description	Avg.	50%	90%
LLM interrupted → LLM generate first text token chunk	478	468	750
First text token chunk → Prefill of speech decoder	15	15	17
Prefill of speech decoder → Generate first speech token chunk	237	235	252
First speech token Chunk → Decode first PCM hunk	11	11	13
Total	745	753	1020

- **Main delay** occurs before the first text token generation (~0.5 s).
- **Speech decoding is extremely fast** (tens of ms).
- Overall, **Freeze-Omni achieves real-time conversational speed (~1.2 s)**, practical for interactive speech dialogue.

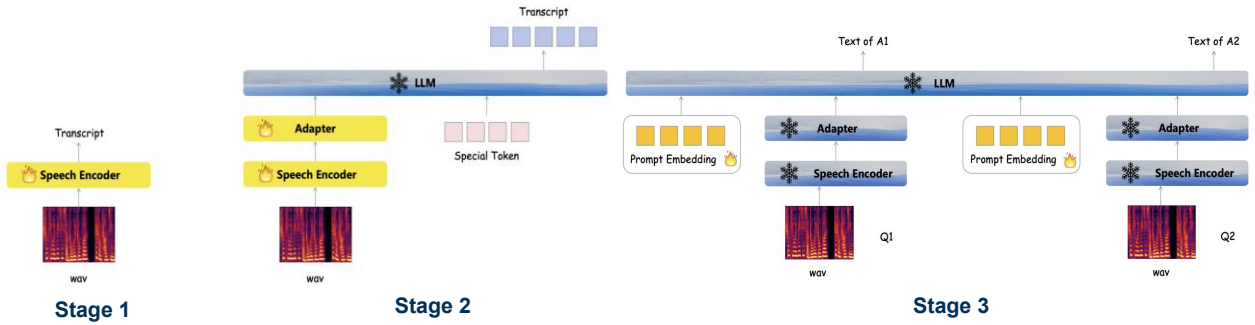
Conclusion

- **Freeze-Omni**:
  a **text–audio multimodal LLM** enabling **low-latency speech-to-speech dialogue** without fine-tuning the LLM.

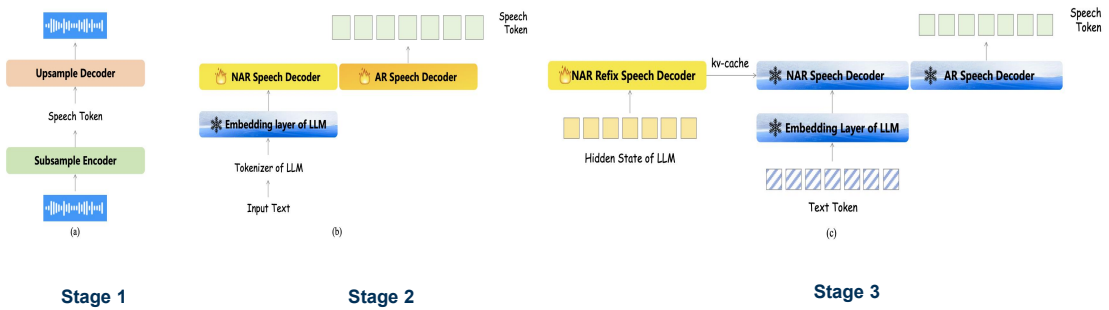
- Achieves **strong performance** across multiple evaluation tasks while keeping the LLM **frozen**.

Provide pathway developing strong performance speech modules to next research

Speech Input



Speech Output

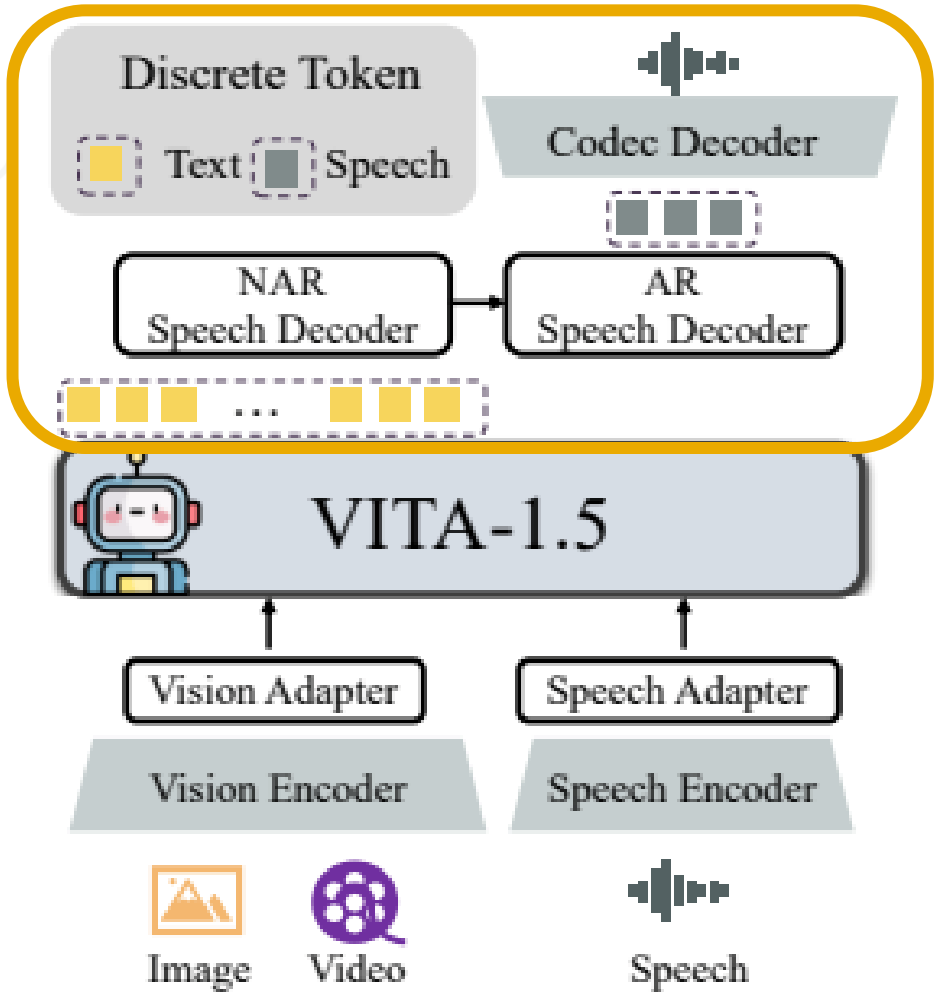


Georgia Tech

History

Fu et al, VITA: Towards Open-Source Interactive Omni Multimodal LLM https://arxiv.org/html/2408.05211v1
Wang et al, Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM https://arxiv.org/html/2411.00774v1
Xie et al, Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities https://arxiv.org/abs/2410.11190
Fu et al, VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction https://arxiv.org/abs/2501.01957

Architecture: Speech Decoders



- **End-to-end speech generation:** replaces external TTS modules (VITA-1.0) with internal decoders + TiCodec for waveform synthesis.
- **Challenge:** LLM natively outputs text tokens only. The new decoders enable output of speech tokens.
- **Two-stage design: (why?)**
- **(1) NAR Decoder:** processes text tokens in parallel and outputs initial speech-token distribution for global semantics
- **(2) AR Decoder:** refines speech tokens step-by-step and improves fluency and naturalness.
- Both decoders: **4 LLaMA-style transformer layers**
- Balances speed (NAR) and quality (AR)
- The output speech tokens are then passed to **TiCodec**, which decodes speech tokens into 24 kHz waveform.

Overview of Model Training

Training Data Overview

- Diverse datasets spanning image, video, text, and speech, in both Chinese and English.
- **Image Captioning:** ShareGPT4V, ALLaVA-Caption etc.
- **Image QA & Reasoning:** LLaVA-150K, LVIS-Instruct etc.
- **OCR & Diagram Understanding:** Anyword-3M etc.
- **Video Tasks:** ShareGemini and synthetic video-QA data
- **Speech Data:** 110k hours speech–transcription, which trains and aligns speech encoder with the LLM. 3k hours text–speech, which trains speech decoder for end-to-end speech synthesis.
- **Text-only Data**

Three-Stage Progressive Training Strategy

- **Challenge:** Modality conflicts as training on speech can degrade visual understanding if done jointly.
- **Solution:** Gradual integration of modalities to preserve previously learned strengths.

Stage 1: Vision-Language Training

Goal: Establish strong alignment between **vision and language** before adding speech.

Stage 1.1 Vision Alignment:

- Train only the **visual adapter** using 20 % of caption data.
- Purpose: let LLM start linking visual features with text space.

Stage 1.2 Vision Understanding:

- Unfreeze **vision encoder + adapter + LLM**.
- Use all caption data to teach descriptive image-to-text generation.

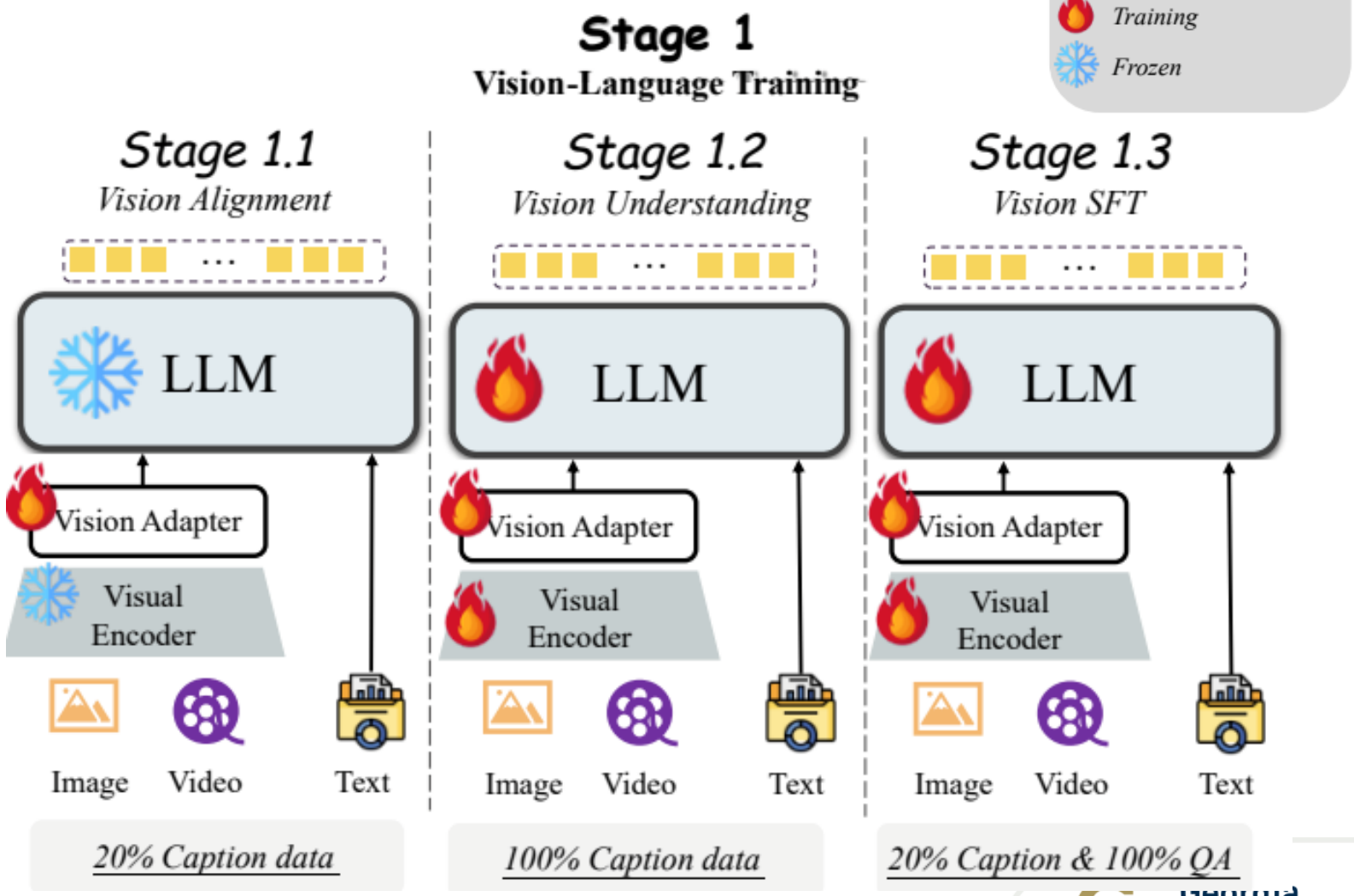
Stage 1.3 Vision SFT (Instruction Tuning):

- Use full **QA datasets + 20 % caption data**.
- Fine-tune all modules to follow instructions and answer visual questions.

Outcome: Model develops rich **visual grounding + instruction-following** ability across images & videos.

LLM: Qwen2-7B

Vision Encoder: InternViT-300M



Stage 2: Audio Input Tuning

Goal: Extend vision-language model to **understand speech** inputs.

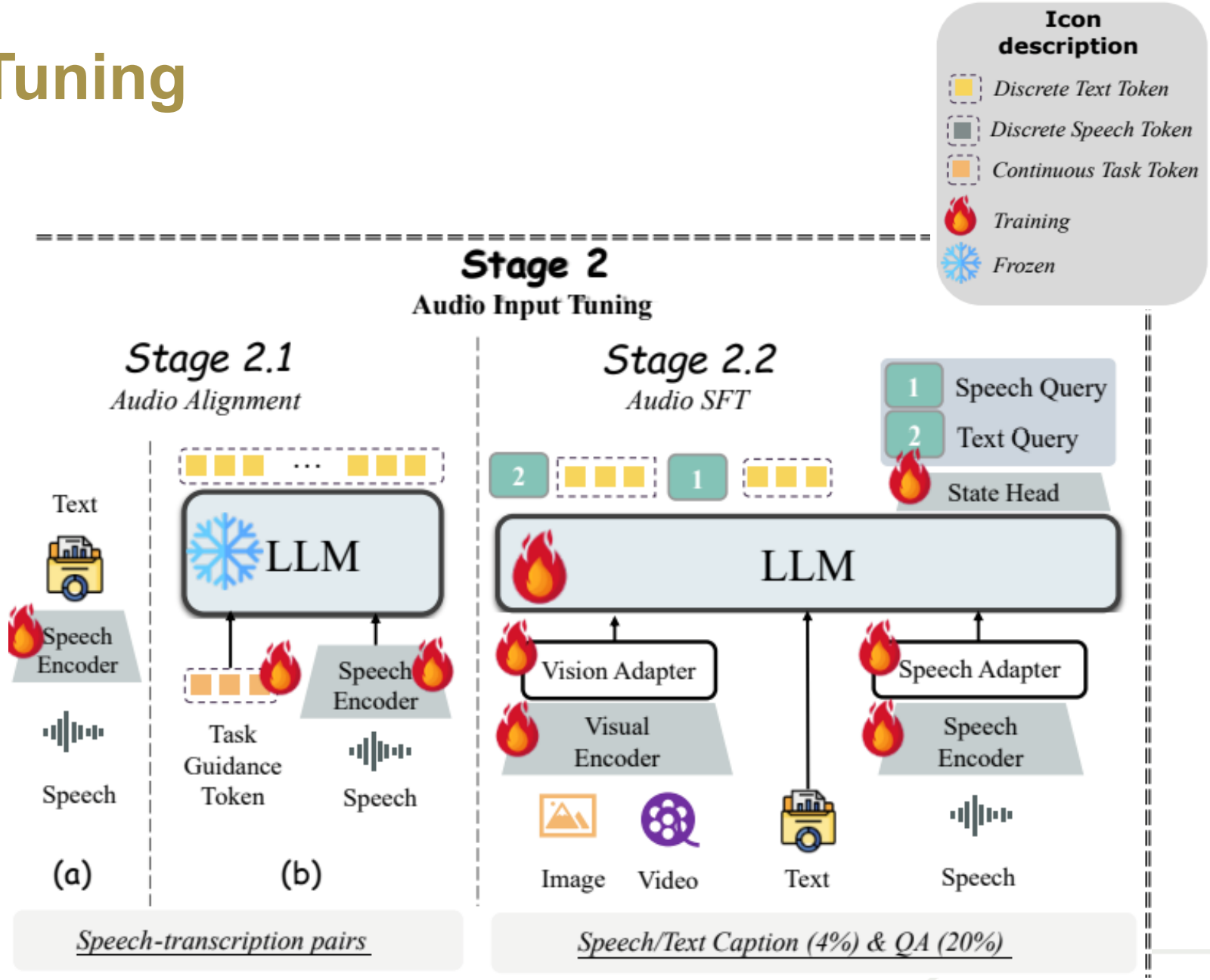
Stage 2.1 Audio Alignment:

- Train on **11 k hrs speech–text pairs**.
- **(a)** Speech encoder: CTC loss for speech-to-text alignment.
- **(b)** Freeze LLM and train **speech adapter + guidance tokens** to feed audio features into LLM.

Stage 2.2 Audio SFT:

- Adds **spoken question-text answer** tasks
- Train all modules (vision + audio encoders/adapters + LLM).
- Add **modality classifier head** to tell speech vs text inputs.

Outcome: Model gains **robust speech comprehension**, enabling real multimodal QA.



Georgia Tech

Stage 3: Audio Output Tuning

Goal: Enable **speech generation** while preserving vision–language and audio-understanding capabilities.

Training data: 3 000 hours of **text–speech pairs**

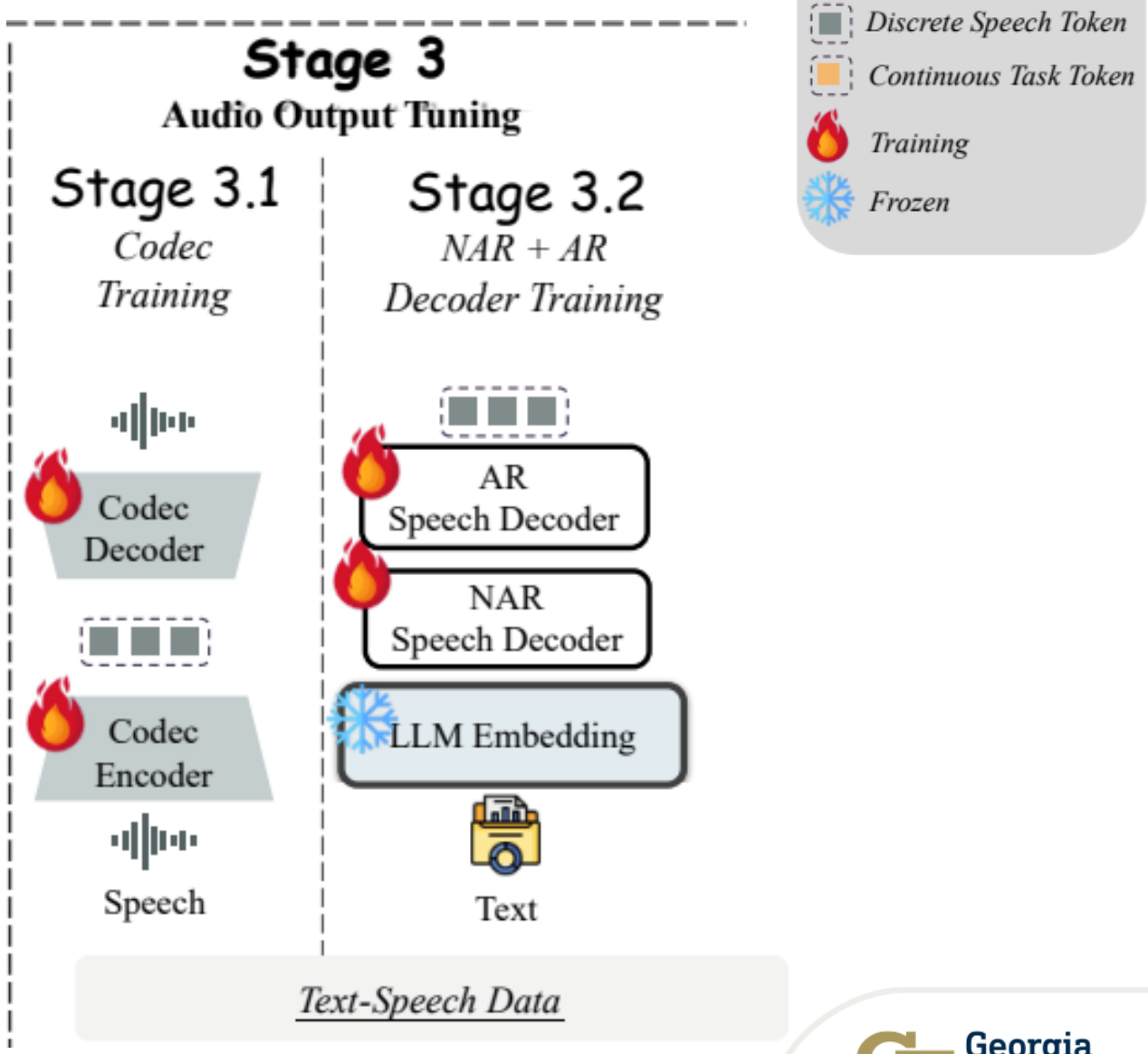
Stage 3.1 Codec Training:

- Train **TiCodec** (single-codebook, 1024 entries).
- Encoder maps waveform to discrete speech tokens; decoder maps tokens back to waveform.
- During inference, only the **codec decoder** is used

Stage 3.2 NAR + AR Decoder Training:

- Freeze LLM weights.
- Use **text embeddings** (from LLM) and **speech tokens** (from codec encoder).
- **NAR decoder:** produces global semantic features in parallel.
- **AR decoder:** predicts high-fidelity speech tokens sequentially.

Outcome: Model gains **end-to-end speech generation** with real-time response and retained multimodal reasoning.



Georgia Tech

Evaluation: Image Understanding

Table 2: **Evaluation on Image Understanding Benchmarks.** VITA-1.5 shows performance comparable to the leading open-source models and advanced closed-source counterparts. MMB refers to MMBench, MMS to MMStar, Hal to HallusionBench, MathV to MathVista, and OCR to OCRBench. Note that after the training of Stages 2 (Audio Input Tuning) and 3 (Audio Output Tuning), VITA-1.5 retains almost its original visual-language capabilities in Stage 1 (Vision-Language Training).

Method	LLM	MMB	MMS	MMMU	MathV	Hal	AI2D	OCR	MMVet	MME	Avg
VILA-1.5	Vicuna-v1.5-13B	68.5	44.2	41.1	42.5	39.3	69.9	460.0	45.0	1718.2	52.1
LLaVA-Next	Yi-34b	77.8	51.6	48.8	40.4	34.8	78.9	574.0	50.7	2006.5	58.3
CogVLM2	Llama3-8B-Instruct	70.7	50.5	42.6	38.6	41.3	73.4	757.0	57.8	1869.5	58.8
InternLM-Xcomposer2	InternLM2-7B	77.6	56.2	41.4	59.5	41.0	81.2	532.0	46.7	2220.4	61.2
Cambrian	Nous-Hermes-2-Yi-34B	77.8	54.2	50.4	50.3	41.6	79.5	591.0	53.2	2049.9	61.4
InternVL-Chat-1.5	InternLM2-20B	79.7	57.1	46.8	54.7	47.4	80.6	720.0	55.4	2189.6	65.1
Ovis1.5	Gemma2-9B-It	77.3	58.1	49.7	65.6	48.2	84.5	752.0	53.8	2125.2	66.9
InternVL2	InternLM2.5-7b	79.4	61.5	51.2	58.3	45.0	83.6	794.0	54.3	2215.1	67.3
MiniCPM-V 2.6	Qwen2-7B	78.0	57.5	49.8	60.6	48.1	82.1	852.0	60.0	2268.7	68.5
Proprietary											
GPT-4V	-	65.5	50.4	59.3	48.2	39.3	71.4	678.0	49.0	1790.3	58.5
GPT-4o mini	-	76.0	54.8	60.0	52.4	46.1	77.8	785.0	66.9	2003.4	66.3
Gemini 1.5 Pro	-	73.9	59.1	60.6	57.7	45.6	79.1	754.0	64.0	2110.6	67.2
GPT-4o	-	82.8	61.6	62.8	56.5	51.7	77.4	663.0	66.5	2328.7	69.3
Claude3.5 Sonnet	-	78.5	62.2	65.9	61.6	49.9	80.2	788.0	66.0	1920.0	69.3
Ours											
VITA-1.0	Mixtral-8x7B	71.8	46.4	47.3	44.9	39.7	73.1	678.0	41.6	2097.0	57.8
VITA-1.5 (Stage 1)	Qwen2-7B	77.1	59.1	53.1	66.2	44.1	80.3	752.0	51.1	2311.0	67.1
VITA-1.5-Audio (Stage 3)	Qwen2-7B	76.7	59.9	52.1	66.2	44.9	79.3	732.0	49.6	2352.0	66.8

Evaluation: Video Understanding

Table 3: **Evaluation on Video Understanding Benchmarks.** Although VITA-1.5 still lags behind models like GPT-4o and Gemini-1.5-Pro, it performs comparably to many open-source models. Note that after the training of Stages 2 (Audio Input Tuning) and 3 (Audio Output Tuning), VITA-1.5 retains almost its original visual-language capabilities in Stage 1 (Vision-Language Training).

Method	LLM	Video-MME w/o sub	Video-MME w/ sub	MVBench	TempCompass
Video-LLaVA	Vicuna-v1.5-13B	39.9	41.6		49.8
SliME	Llama3-8B-Instruct	45.3	47.2	-	-
LongVA	Qwen2-7B	52.6	54.3	-	57.0
VILA-1.5	Llama3-8B-Instruct	-	-	-	58.8
InternLM-XComposer-2.5	InternLM2-7B	-	-	-	62.1
LLaVA-OneVision	Qwen2-7B	58.2	61.5	56.7	64.2
InternVL-2	InternLM2.5-7b	-	-	-	66.0
MiniCPM-V-2.6	Qwen2-7B	60.9	63.7	-	66.3
Proprietary					
GPT-4o-mini	-	64.8	68.9	-	
Gemini-1.5-Pro	-	75.0	81.3	-	67.1
GPT-4o	-	71.9	77.2	-	73.8
Ours					
VITA-1.0	Mixtral-8x7B	55.8	59.2	-	62.3
VITA-1.5 (Stage 1)	Qwen2-7B	56.8	59.5	56.8	65.5
VITA-1.5 (Stage 3)	Qwen2-7B	56.1	58.7	55.4	66.7

Evaluation: Automatic Speech Recognition (ASR)

Table 4: **Evaluation on ASR Benchmarks.** VITA-1.5 has demonstrated strong performance in both Mandarin and English ASR tasks. It outperforms specialized speech models, achieving better results in both languages.

Model	CN (CER↓)			Eng (WER↓)			
	aishell-1	test net	test meeting	dev clean	dev other	test clean	test other
Wav2vec2-base	-	-	-	6.0	13.4	-	-
Mini-Omini2	-	-	-	4.8	9.8	4.7	9.4
Freeze-Omini	2.8	12.6	14.2	4.2	10.2	4.1	10.5
Ours							
VITA-1.0	-	12.2	16.5	7.6	16.6	8.1	18.4
VITA-1.5	**2.2**	**8.4**	**10.0**	**3.3**	**7.2**	**3.4**	**7.5**

Conclusion: Strengths & Limitations

Strengths

- **Unified multimodal integration:** Seamlessly aligns vision, language, and speech through progressive three-stage training.
- **End-to-end speech interaction:** Supports real-time speech input and output without external ASR/TTS modules.
- **Strong visual reasoning:** Maintains relatively strong image and video understanding even after audio integration.
- **Cross-lingual speech understanding:** Demonstrates robust ASR in both English and Mandarin, showing language-agnostic generalization.
- **Efficient modular design:** Modality-specific adapters and frozen LLM backbone prevent catastrophic forgetting during multi-stage tuning.

Weaknesses / Limitations

- **Video reasoning gap:** Performance lags behind proprietary models due to coarse frame sampling and limited temporal modeling.
- **Synthetic speech supervision:** Reliance on TTS-generated speech data limits diversity and naturalness.
- **Limited data transparency:** Internal speech corpora are not publicly available, affecting reproducibility.
- **Scalability ceiling:** Model size and compute budget constrain its ability to match closed-source models trained on multi-million-hour datasets.