

# Multimodal model

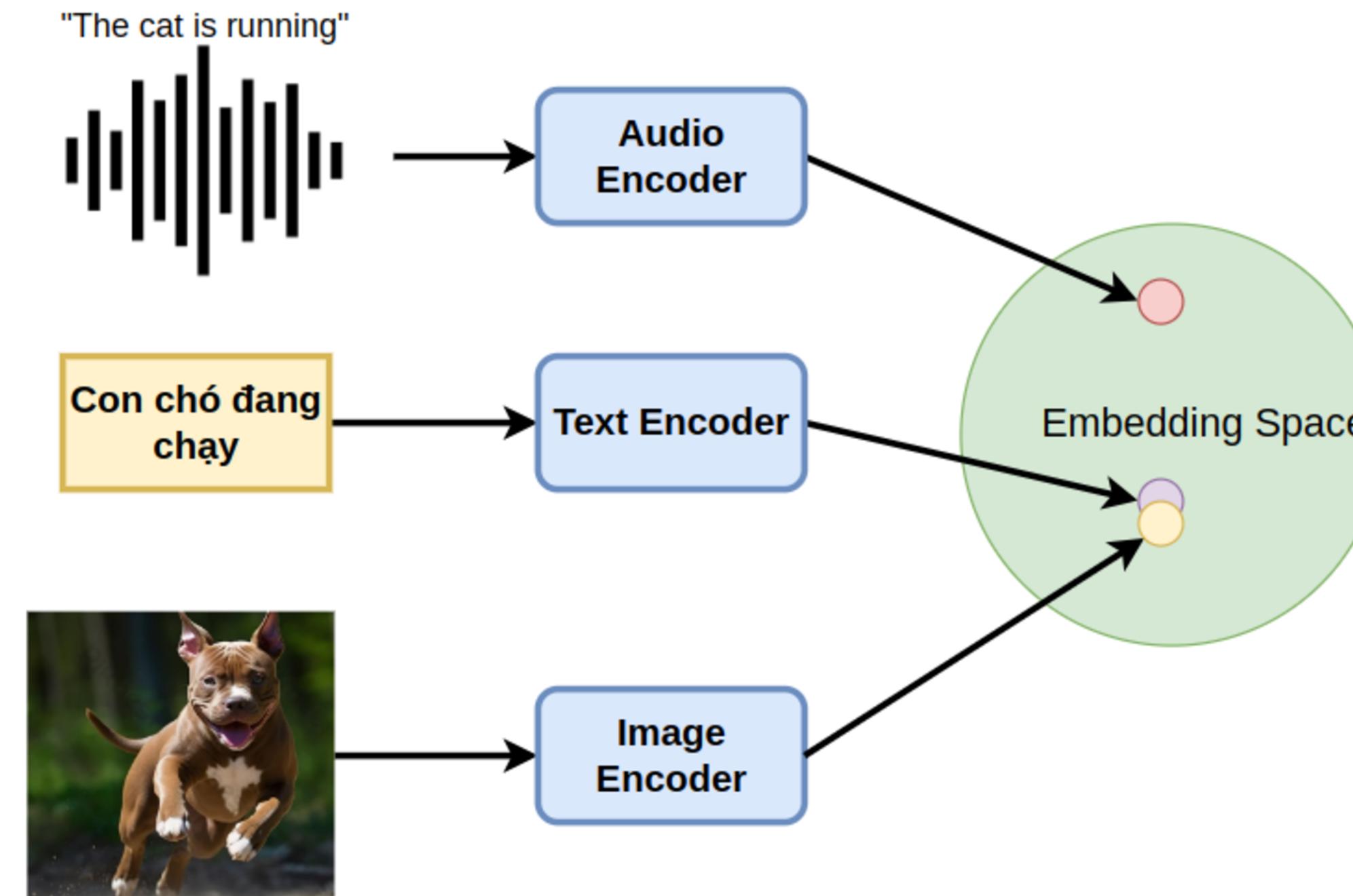
Khanhnd10 (Aka khanhnd65)

# Table of content

- 1.Text-Speech bottleneck embedding
- 2.Text-Speech aligned representations
- 3.Downstream multimodal model

# I. Text-Speech bottleneck embedding

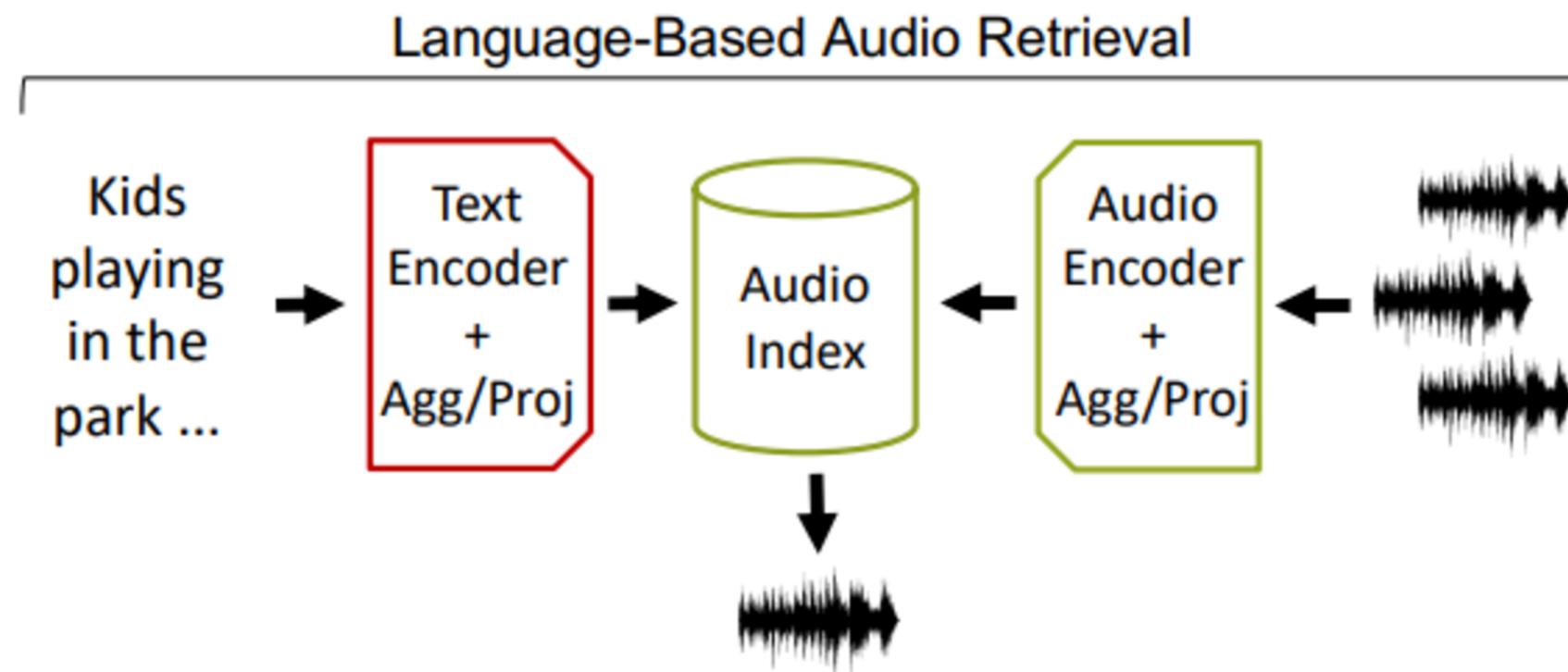
## 1. Overview



# I. Text-Speech bottleneck embedding

## 2. Applications

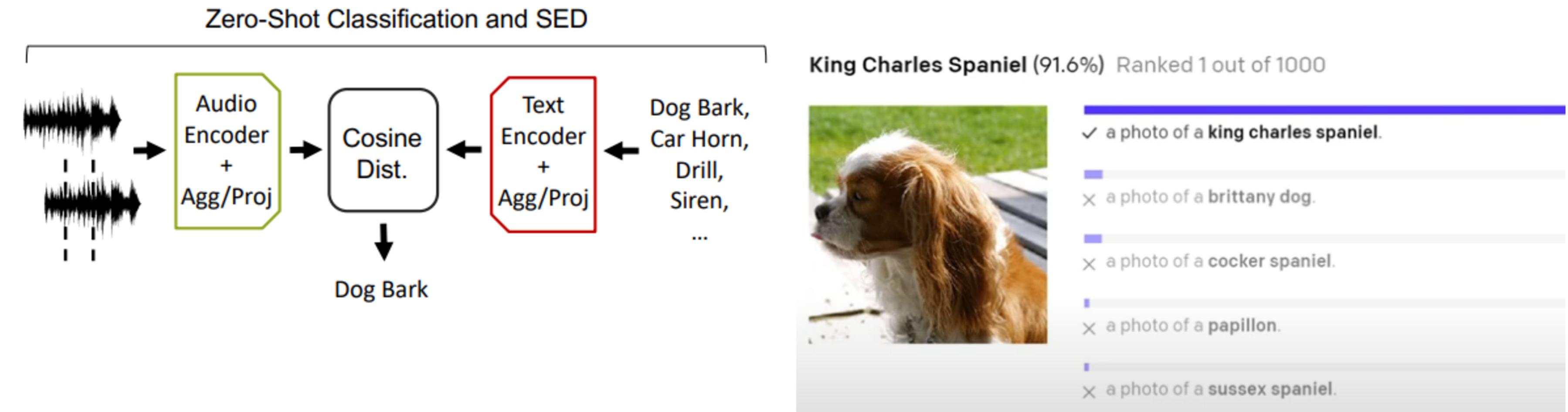
### a) Cross model audio-text retrieval



# I. Text-Speech bottleneck embedding

## 2. Applications

### b) Zero shot downstream task



# I. Text-Speech bottleneck embedding

## 2. Applications

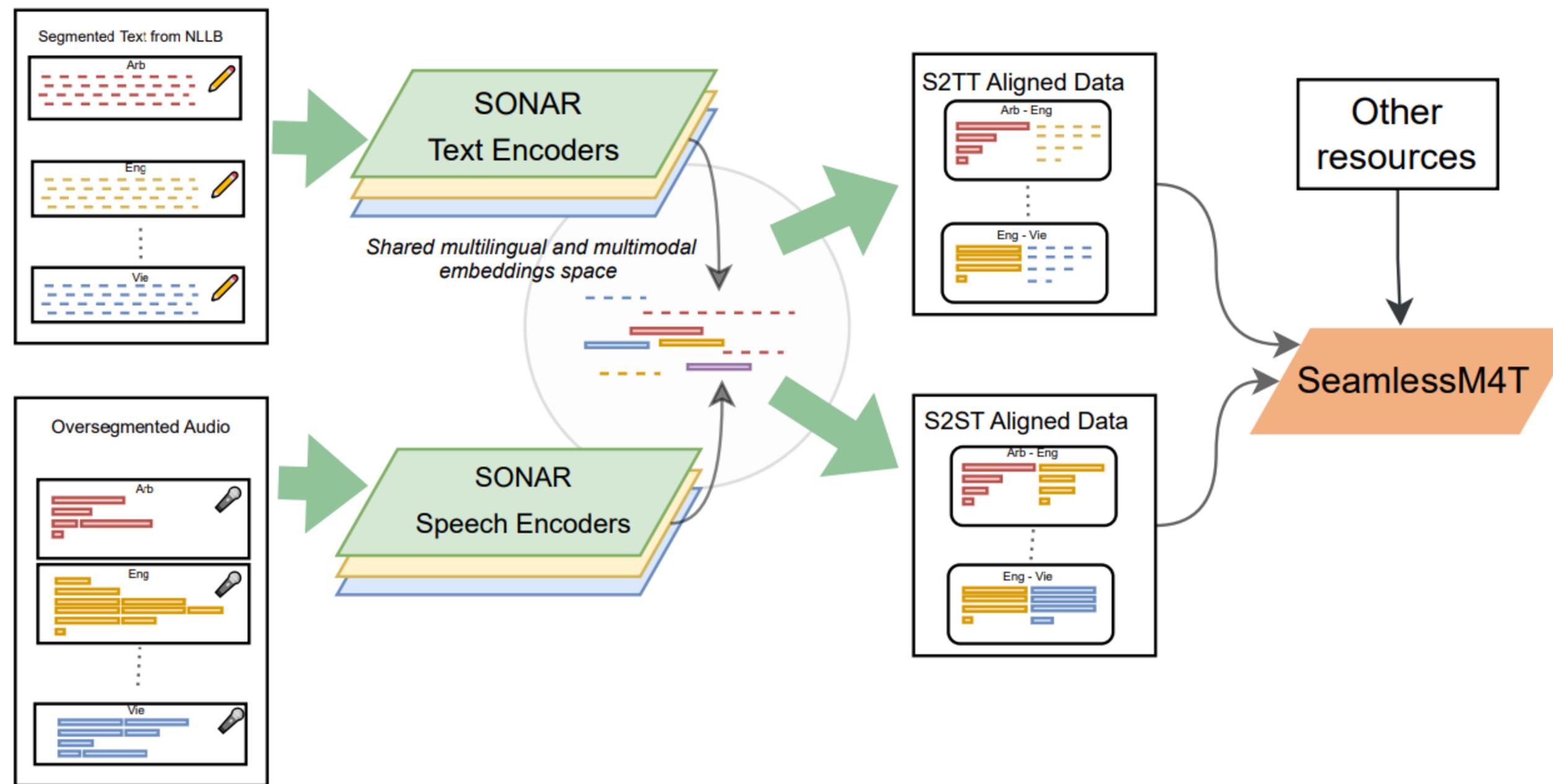
### b) Zero shot downstream task

tinued improvement, we estimate around a 1000x increase in compute is required for zero-shot CLIP to reach overall state-of-the-art performance. This is infeasible to train with current hardware. Further research into improving upon the computational and data efficiency of CLIP will be necessary.

# I. Text-Speech bottleneck embedding

## 2. Applications

### c) Data Mining

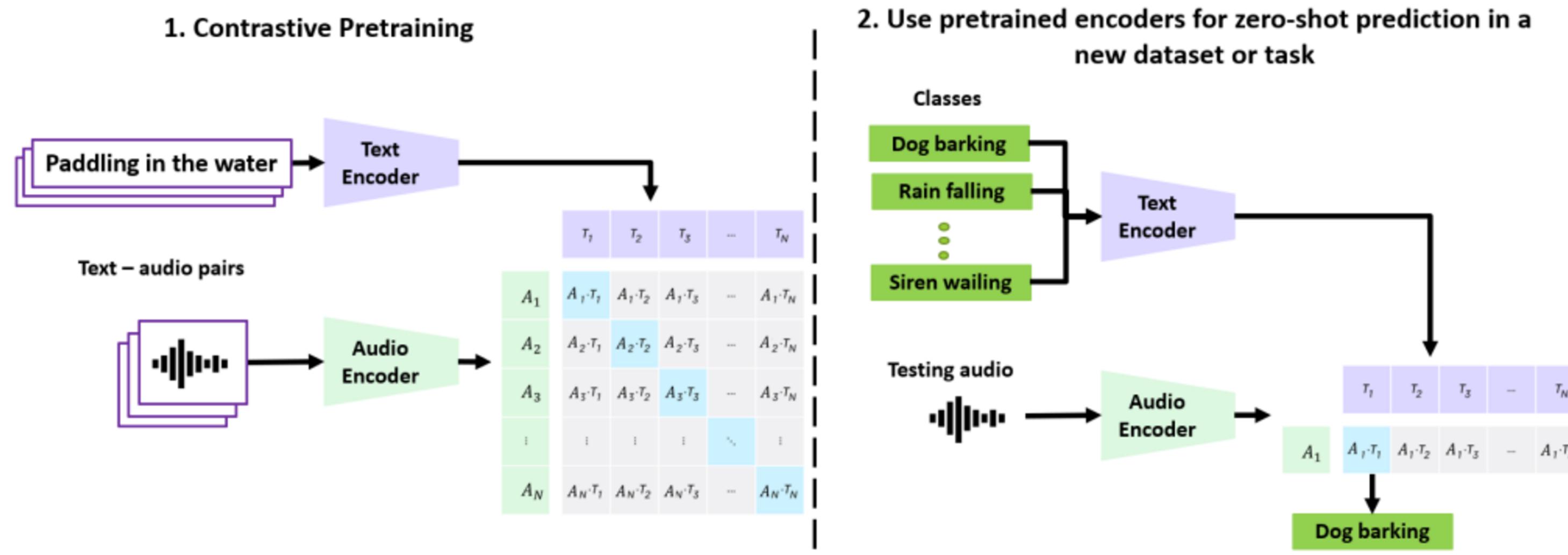


**Figure 2:** Workflow of the SONAR encoding and mining processes.

# I. Text-Speech bottleneck embedding

## 2. Approaches

### a) CLAP (Microsoft)



# I. Text-Speech bottleneck embedding

## 2. Approaches

### a) SONAR (Meta - 2023)

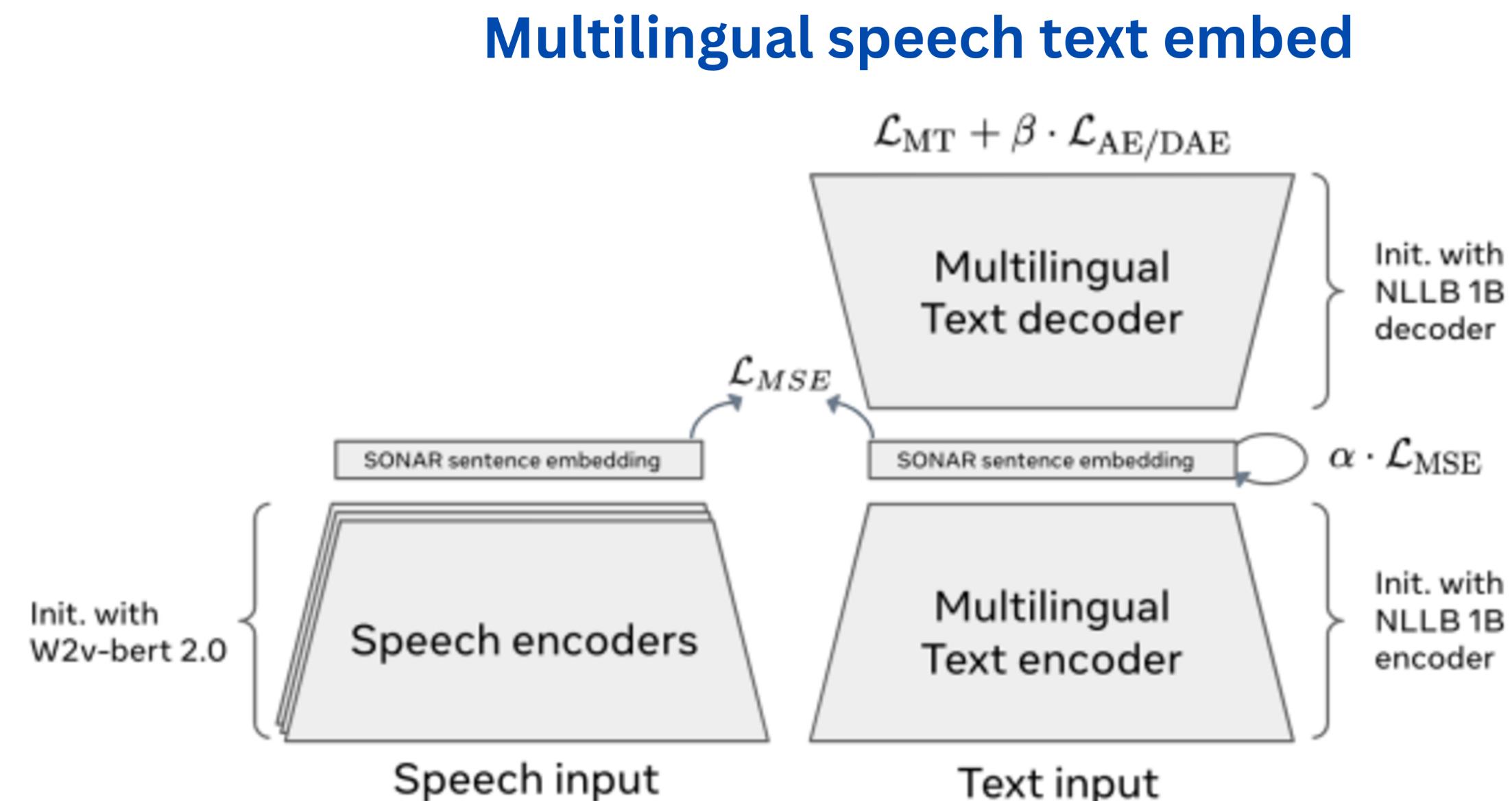


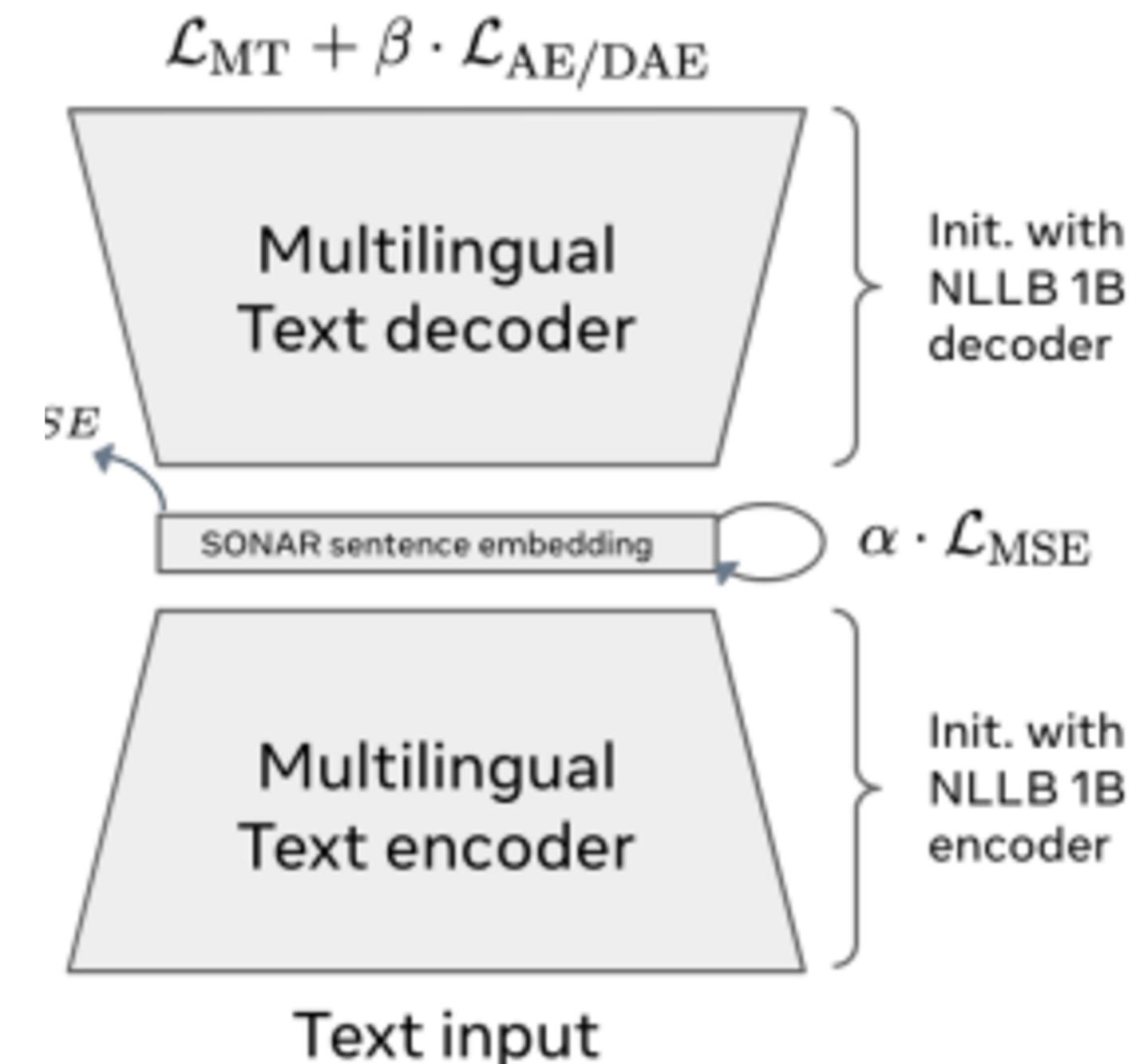
Figure 1: SONAR architecture.

# I. Text-Speech bottleneck embedding

## 2. Approaches

### a) SONAR (Meta - 2023)

First, we achieve multilingual text embedding vector by training a Machine translation model

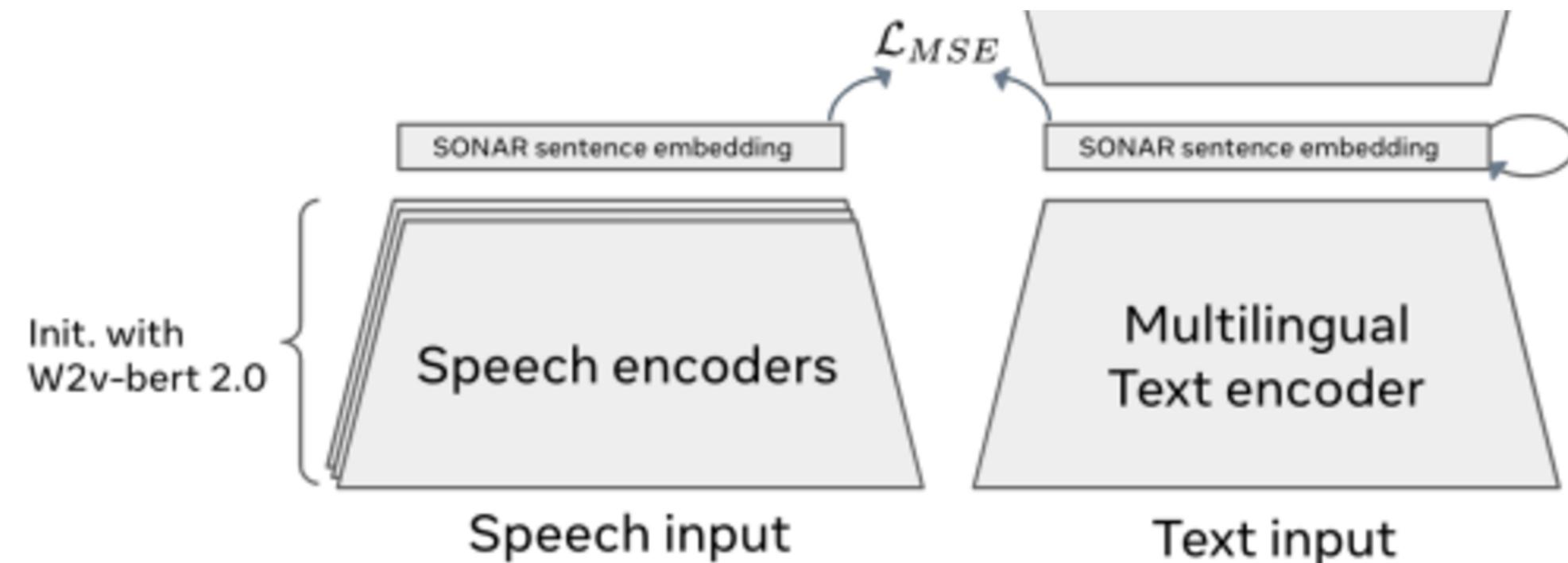


# I. Text-Speech bottleneck embedding

## 2. Approaches

### a) SONAR (Meta - 2023)

Then, they train a speech encoder that can map to the latent space of the text encoder using MSE loss



# II. Text-Speech aligned representations

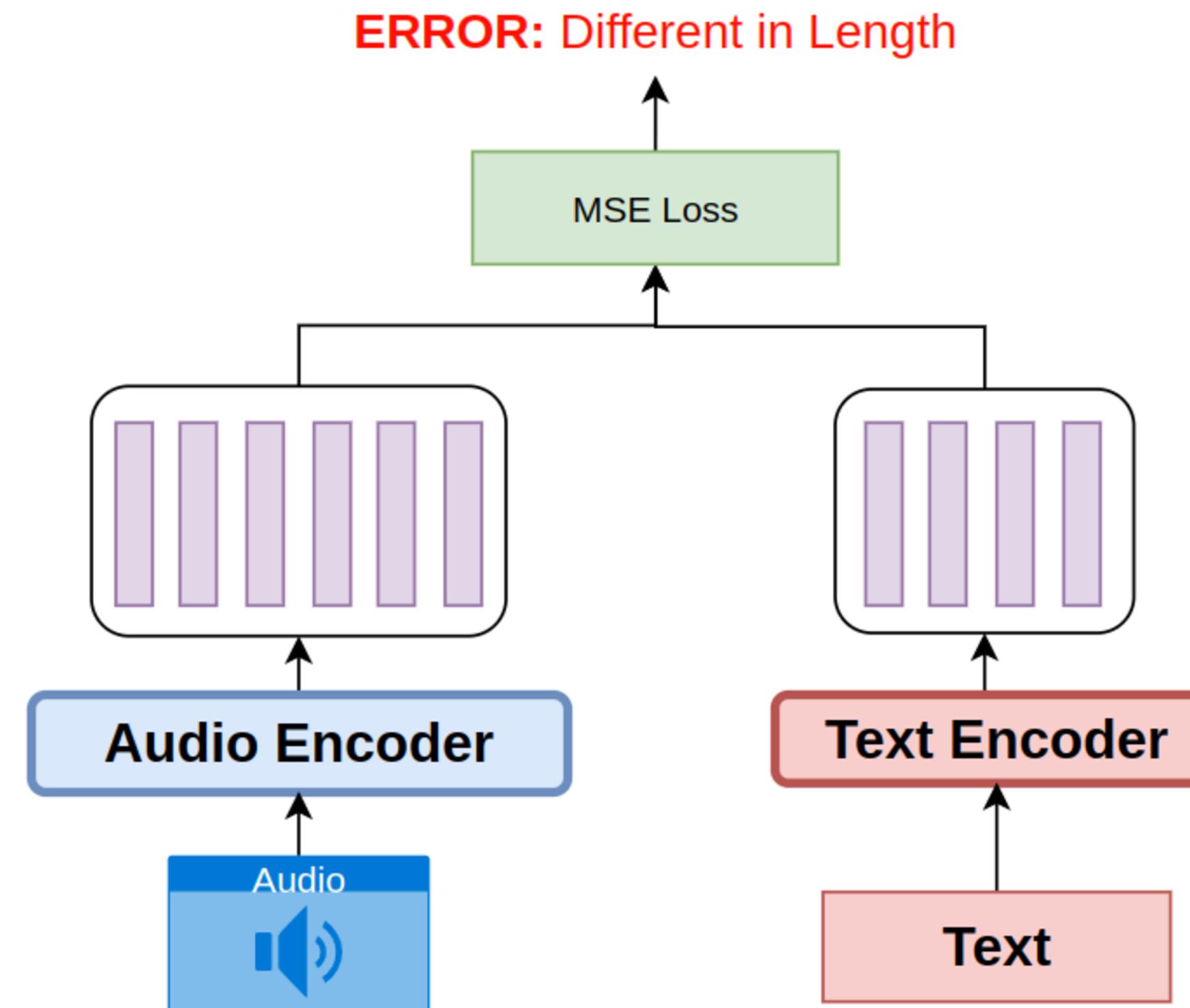
## 1. Motivation

- Self-supervised model is fine-tuned to the task of interest
- Require the learned representations to capture ***acoustic, prosodic, speaker, and linguistic*** characteristics as well as the ***semantics*** of the speech.

-> **Joint pre-training** from both **speech** and **text** modalities is a natural extension for improved generalization

# II. Text-Speech aligned representations

## 2. Challenge



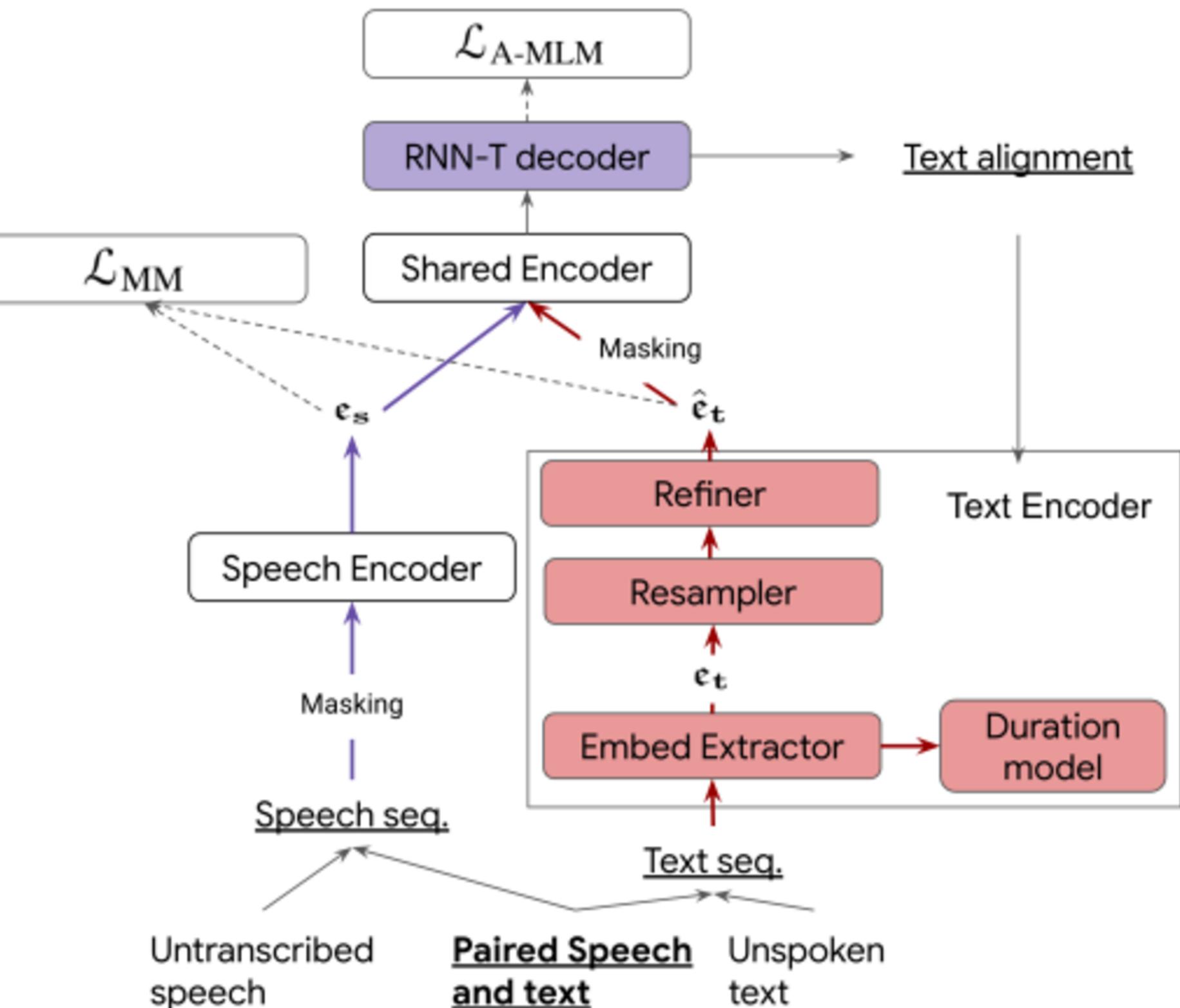
# II. Text-Speech aligned representations

## 3. Approaches

### a) MAESTRO (Google - 2022)

Table 1: *Multi-domain ASR: SpeechStew results on 5 domains.*  
All the models use 0.6B parameters.

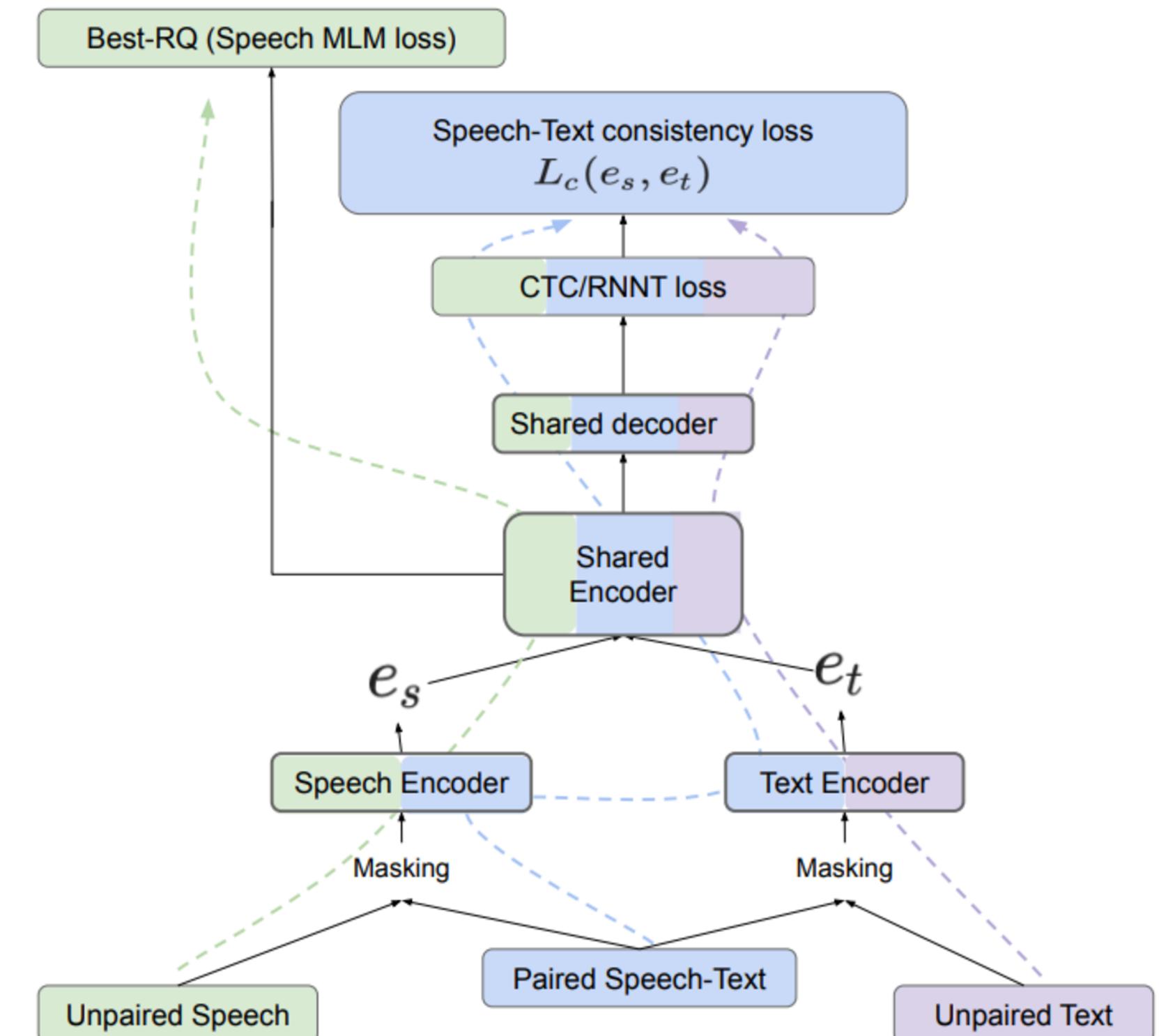
| Method                 | LS-test    |       | AMI        |             | TED        |            | SWB      |            | CV             |
|------------------------|------------|-------|------------|-------------|------------|------------|----------|------------|----------------|
|                        | clean      | other | ihm        | sdm         | swb        | chm        |          |            |                |
| Wav2vec2               | 1.7        | 3.3   | 9.6        | 23.8        | 5.7        | 4.9        | 10.8     | 8.5        |                |
| W2v-BERT<br>+ LM       | 1.6        | 3     | 9.1        | 23.1        | 5.4        | 4.5        | 9        | 8.6        | <b>1.5</b> 2.8 |
| SLAM                   | 1.6        | 3.1   | 9.3        | 23.5        | 5.6        | 4.6        | 9.1      | 8.6        |                |
| TTS4Pretrain2          | 1.6        | 2.8   | 8.7        | <b>21.9</b> | 5          | 4.5        | 8.5      | 8.4        |                |
| <b>Maestro</b><br>+ LM | <b>1.5</b> | 2.8   | <b>8.5</b> | <b>21.9</b> | <b>4.9</b> | <b>4.3</b> | <b>8</b> | <b>8.1</b> | <b>1.5</b> 2.7 |



# II. Text-Speech aligned representations

## 3. Approaches

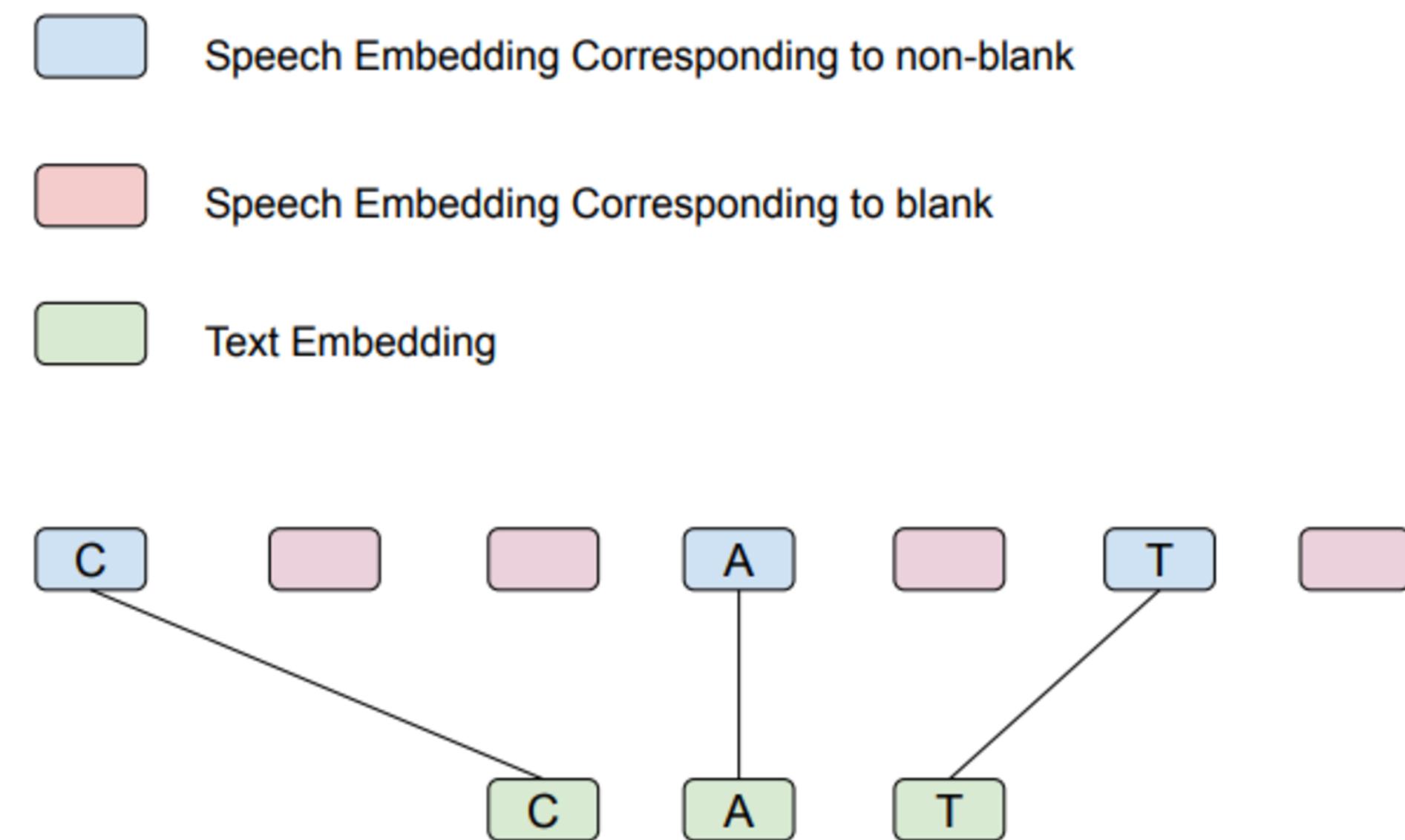
### b) ASTRA (Google - 2024)



# II. Text-Speech aligned representations

## 3. Approaches

### b) ASTRA (Google - 2024)



# II. Text-Speech aligned representations

## 3. Approaches

### b) ASTRA (Google - 2024)

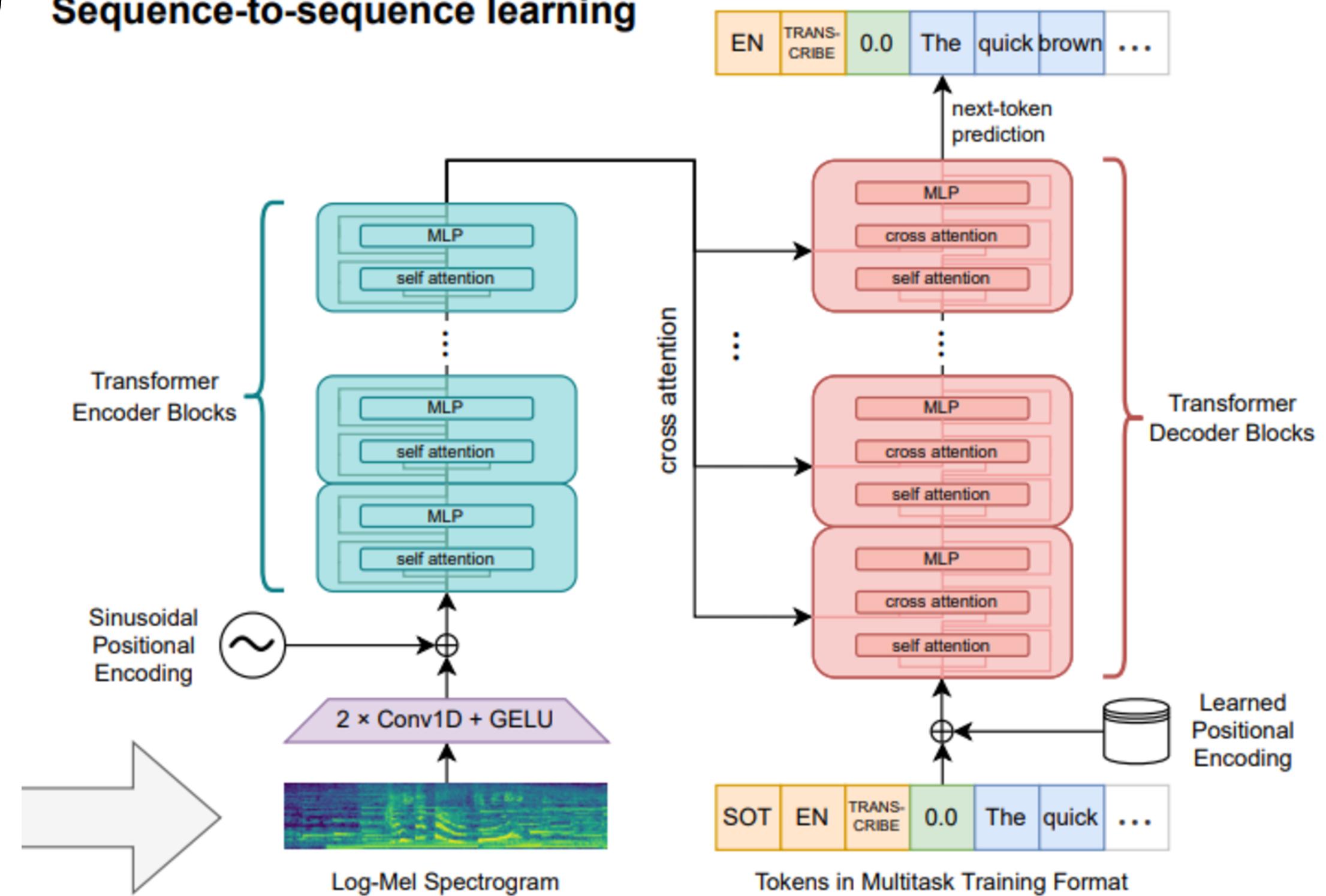
| Model                                 | Pretraining data                   | # params | average CER |
|---------------------------------------|------------------------------------|----------|-------------|
| Vanilla Conformer                     | YT-56-U                            | 300M     | 13.04       |
| w2v-bert-51 [24]                      | VoxPopuli, MLS, CommonVoice, BABEL | 600M     | 14.1        |
| mSLAM [24]                            | VoxPopuli, MLS, CommonVoice, BABEL | 600M     | 14.6        |
| Text injection + duration model [2]   | YT-56-U                            | 300M     | 13.27       |
| Text injection + duration model + VAE | YT-56-U                            | 300M     | 12.38       |
| ASTRA                                 | YT-56-U                            | 300M     | 12.38       |

# III. Downstream multimodal model

## 1. Multi-task model

### a) Whisper (OpenAI - 2022)

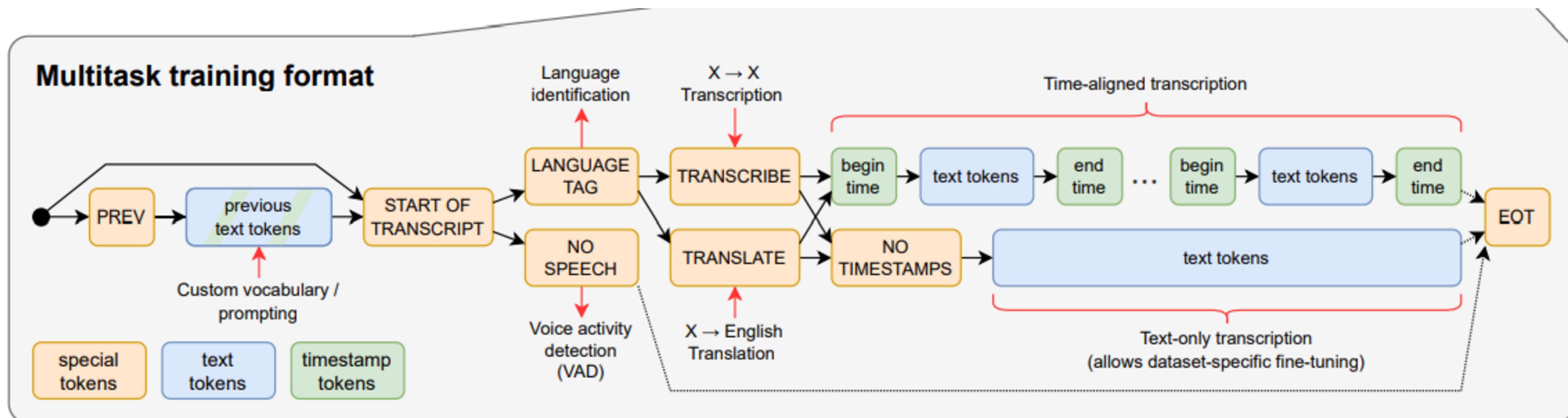
Sequence-to-sequence learning



# III. Downstream multimodal model

## 1. Multi-task model

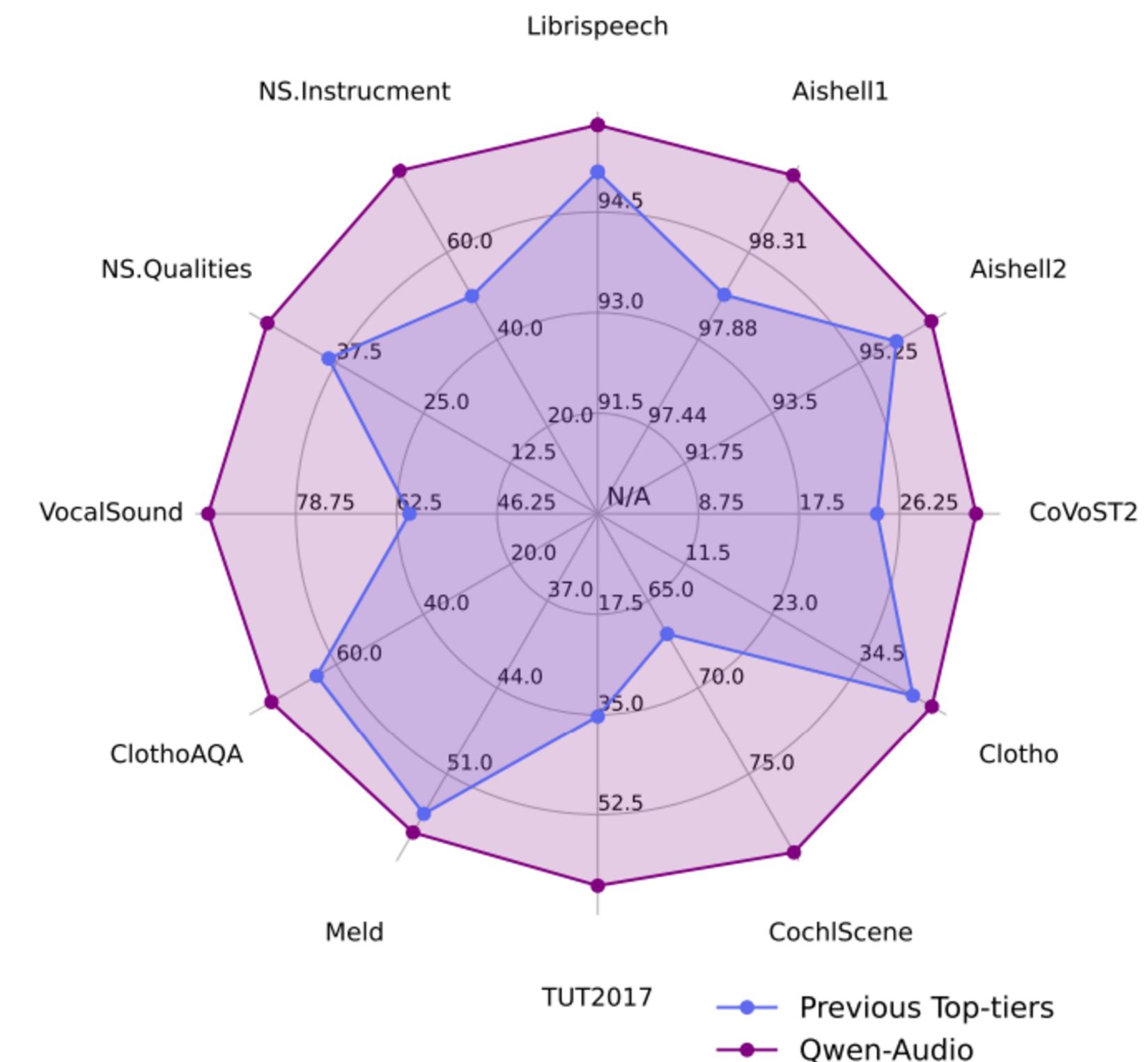
### a) Whisper (OpenAI - 2022)



# III. Downstream multimodal model

## 1. Multi-task model

### b) Qwen-Audio (Alibaba -2023)



# III. Downstream multimodal model

## 1. Multi-task model

### b) Qwen-Audio (Alibaba -2023)

The interface shows three examples of speech recognition and translation:

- recognize the speech, and translate it into English  
OK. The sentence “大家好我们来讨论一下公司周末团建的地点吧大家有什么意见呢” translated into English is “Hello, let's discuss the location for our company's weekend team building. What are your opinions?”
- recognize the speech, and translate it into English  
OK. The sentence “I think we can go to the countryside and find a beautiful natural area.” translated into English is “I think we can go to the countryside and find a beautiful natural area.”
- recognize the speech, and translate it into English  
OK. The sentence “わかりました。田舎に行くことに同意します。田舎暮らしを体験してみるのもいいかもしれません。” translated into English is “I understand. I agree to go to the countryside. It's also good to experience country life.”

(a) Multilingual ASR and Translation

The interface shows two examples of audio analysis:

- What are the emotions in audio 1 and audio 2, respectively?  
The tone of the woman in audio 1 is cheerful, possibly talking about something pleasant. The tone of the woman in audio 2 is fearful, possibly discussing something terrifying.
- How to respond to the questions in audio 1 and audio 2?  
In audio 1, when the woman asks if you're okay, you can respond with, “Thank you for your concern, I'm fine now, just felt a bit uncomfortable all of a sudden.” In audio 2, considering the woman's fearful tone and discussing something terrifying, you can respond to her with either, “I'm scared, I don't know what to do” or “I'm feeling really upset, I don't know how to face this situation.”

(b) Multiple Audio Analysis

The interface shows one example of sound understanding and reasoning:

Recognize the sound information and assume that the sound is happening around the user. Please provide the user with some handling suggestions.

The sound is of breaking glass. Here are some suggestions for handling the situation:

- Turn off the source of the sound: If the sound is coming from a broken glass item, turn off the source of the sound as soon as possible to prevent further damage.
- Clean up the broken glass: Use gloves and a broom or vacuum cleaner to clean up the broken glass. Be careful not to cut yourself.
- Assess the damage: If the broken glass is in a location that is not easily accessible or if it poses a safety hazard, contact a professional for assistance.
- Cover the area: Cover the area where the broken glass is located to prevent injury or further damage.
- Avoid walking on broken glass: If the broken glass is on a floor, do not walk on it. Instead, use a broom or vacuum cleaner to clean it up or cover it with a piece of cardboard or tape.

(c) Sound Understanding and Reasoning

# III. Downstream multimodal model

## 1. Multi-task model

### b) Qwen-Audio (Alibaba -2023)

A user interface showing a man icon, a waveform icon, and a bird image. A blue box contains the text: "Write a poem about the above audio". Below is a poem:

Birds are singing in the trees,  
A gentle breeze is blowing free,  
The rustling leaves are music,  
As nature's symphony goes on.  
The sun is shining bright and warm,  
A peaceful atmosphere is sown,  
In the midst of nature's beauty,  
We can find serenity and duty.  
The birds take flight with ease,  
Their chirping voices fill the trees,  
A symphony of nature's grace,  
A moment to pause and embrace.  
The wind whispers through the branches,  
As the leaves dance in the branches,  
A peaceful moment in time,  
A chance to let go and unwind.  
So let us take a moment here,  
To appreciate the beauty so dear,  
In the midst of nature's wonder,  
We can find a sense of wonder.

At the bottom right is a purple SWRT logo.

(d) Audio-Motivated Creative Writing

A user interface showing a woman icon, a waveform icon, and a piano image. A blue box contains the text: "what is it". Below is the response: "A solo piano piece." with a purple SWRT logo.

The woman then asks: "Describe this music in detail". Below is the detailed description: "This is a solo piano piece. It is fast tempo with a playful and humorous mood. The piece is in the key of C major and is in 4/4 time signature. The melody is played on the right hand with the left hand playing chords. The piece is a classic ragtime style." with a purple SWRT logo.

The woman then asks: "Give me some recommendations for similar music.". Below is the list of recommendations: "some similar music to "Solo Piano Ragtime" could be:  
1. "The Entertainer" by Scott Joplin  
2. "Maple Leaf Rag" by Scott Joplin  
3. "Ave Maria" by Franz Schubert  
4. "Für Elise" by Ludwig van Beethoven  
5. "The Blue Danube" by Johann Strauss II" with a purple SWRT logo.

(e) Music Appreciation

A user interface showing a woman icon, a waveform icon, and a waveform image. A blue box contains the text: "change "joyful" to "interesting"". Below is the edited audio with a purple SWRT logo.

The woman then asks: "change the person name to "John"". Below is the edited audio with a purple SWRT logo.

The woman then asks: "Edit out the word "what", so I can turn it into a meme". Below is the edited audio with a purple SWRT logo.

Below is the note: "The word "what" starts at 3.04 seconds and ends at 3.23 seconds." with a purple SWRT logo.

(f) Speech Editing with SWRT

# III. Downstream multimodal model

## 1. Multi-task model

### b) Qwen-Audio (Alibaba -2023)

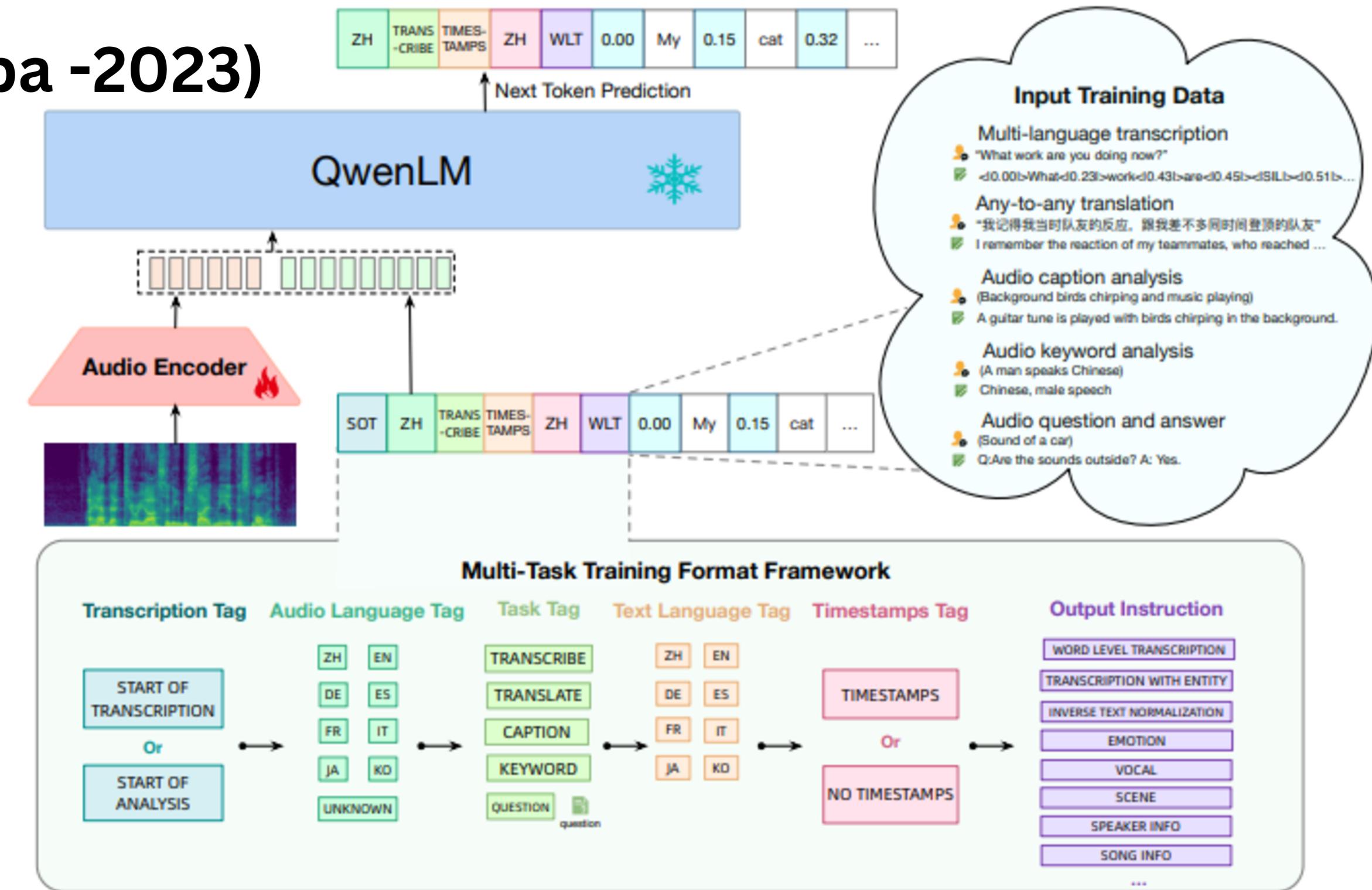


Figure 3: The overview of Qwen-Audio architecture and multitask-pretraining.

# III. Downstream multimodal model

## 1. Multi-task model

### b) Qwen-Audio (Alibaba -2023)

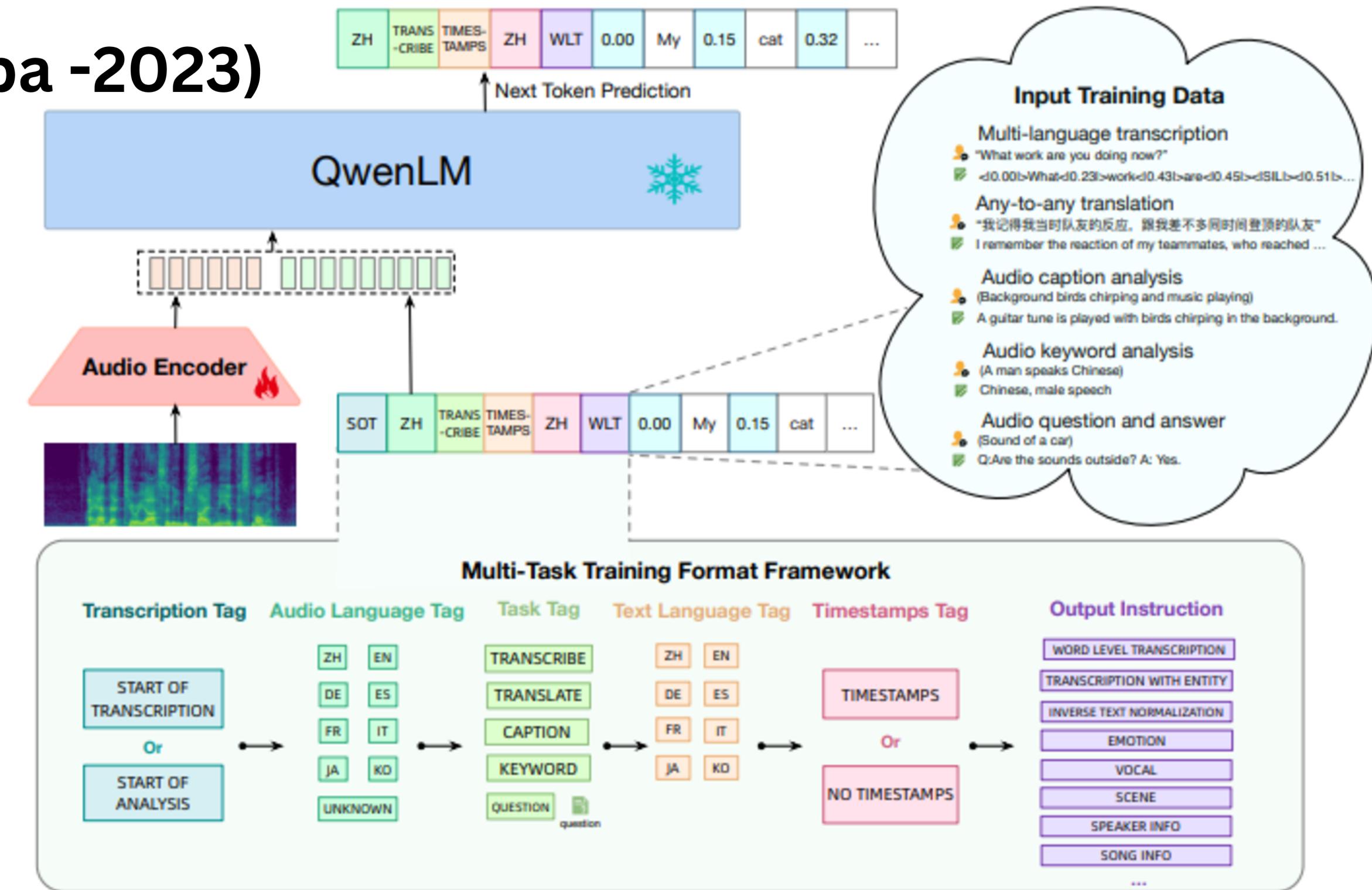


Figure 3: The overview of Qwen-Audio architecture and multitask-pretraining.

# III. Downstream multimodal model

## 1. Multitask model

### b) Qwen-Audio (Alibaba -2023)

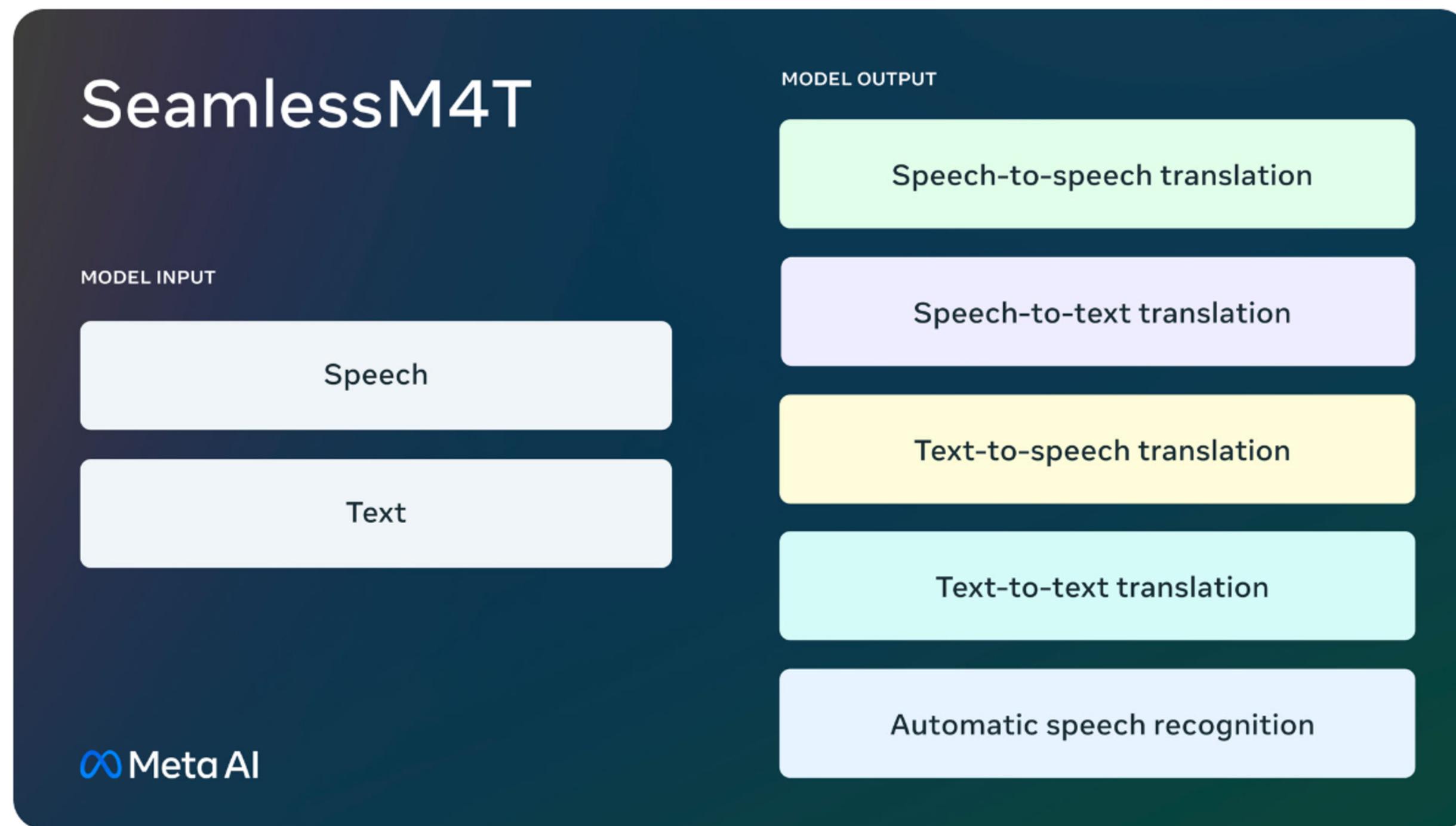
Table 1: Multi-task pre-training dataset.

| Types      | Task        | Description  | Hours |
|------------|-------------|--|-------|
| Speech     | ASR         | Automatic speech recognition (multiple languages)      | 30k   |
|            | S2TT        | Speech-to-text translation                             | 3.7k  |
|            | OSR         | Overlapped speech recognition                          | <1k   |
|            | Dialect ASR | Automatic dialect speech recognition                   | 2k    |
|            | SRWT        | English speech recognition with word-level timestamps  | 10k   |
|            |             | Mandarin speech recognition with word-level timestamps | 11k   |
|            | DID         | Dialect identification                                 | 2k    |
|            | LID         | Spoken language identification                         | 11.7k |
|            | SGC         | Speaker gender recognition (biologically)              | 4.8k  |
|            | ER          | Emotion recognition                                    | <1k   |
|            | SV          | Speaker verification                                   | 1.2k  |
|            | SD          | Speaker diarization                                    | <1k   |
|            | SER         | Speech entity recognition                              | <1k   |
| Sound      | KS          | Keyword spotting                                       | <1k   |
|            | IC          | Intent classification                                  | <1k   |
|            | SF          | Slot filling   | <1k   |
|            | SAP         | Speaker age prediction                                 | 4.8k  |
|            | VSC         | Vocal sound classification                             | <1k   |
| Music&Song | AAC         | Automatic audio caption                                | 8.4k  |
|            | SEC         | Sound event classification                             | 5.4k  |
|            | ASC         | Acoustic scene classification                          | <1k   |
|            | SED         | Sound event detection with timestamps                  | <1k   |
|            | AQA         | Audio question answering                               | <1k   |
| Music&Song | SID         | Singer identification                                  | <1k   |
|            | SMER        | Singer and music emotion recognition                   | <1k   |
|            | MC          | Music caption  | 25k   |
|            | MIC         | Music instruments classification                       | <1k   |
|            | MNA         | Music note analysis such as pitch, velocity            | <1k   |
|            | MGR         | Music genre recognition                                | 9.5k  |
|            | MR          | Music recognition                                      | <1k   |
|            | MQA         | Music question answering                               | <1k   |

# III. Downstream multimodal model

## 1. Multitask model

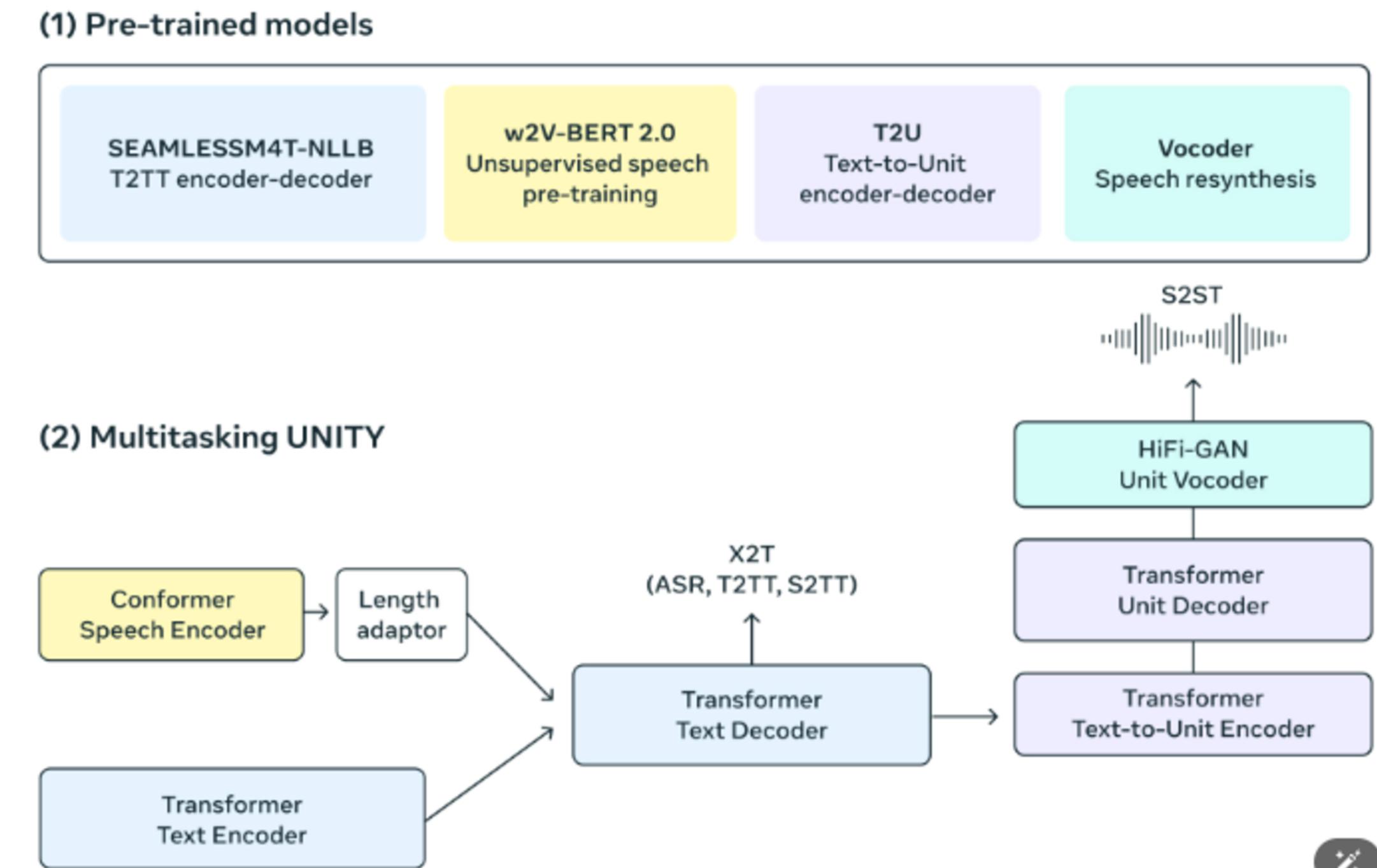
### c) SeamlessM4T (Meta -2023)



# III. Downstream multimodal model

## 1. Multitask model

### c) SeamlessM4T (Meta -2023)

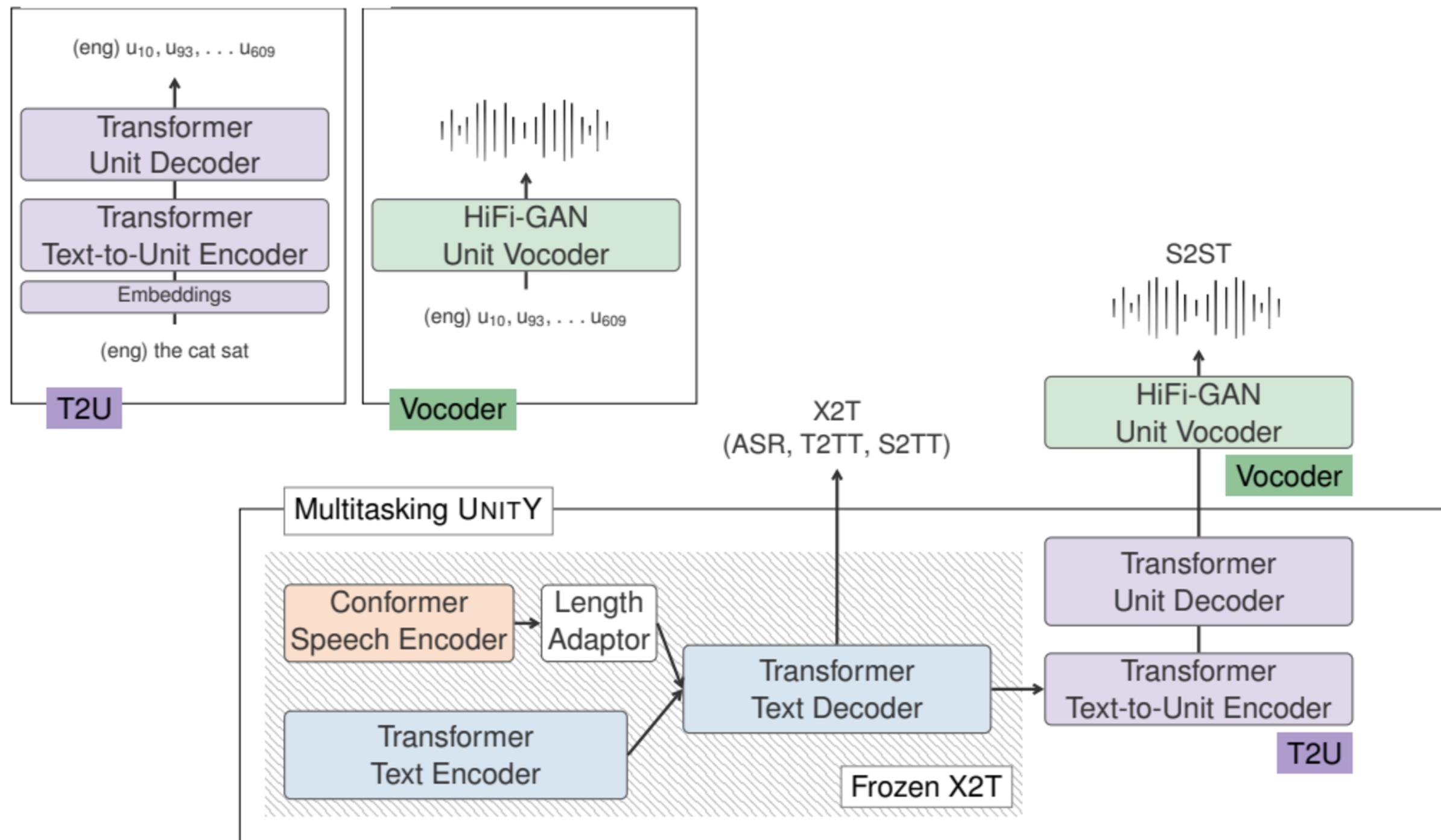


# III. Downstream multimodal model

## 1. Multitask model

### c) SeamlessM4T (Meta -2023)

(1) Pre-trained models



# III. Downstream multimodal model

## 2. Speech LLM

### a) MINT (Bytedance -2024)

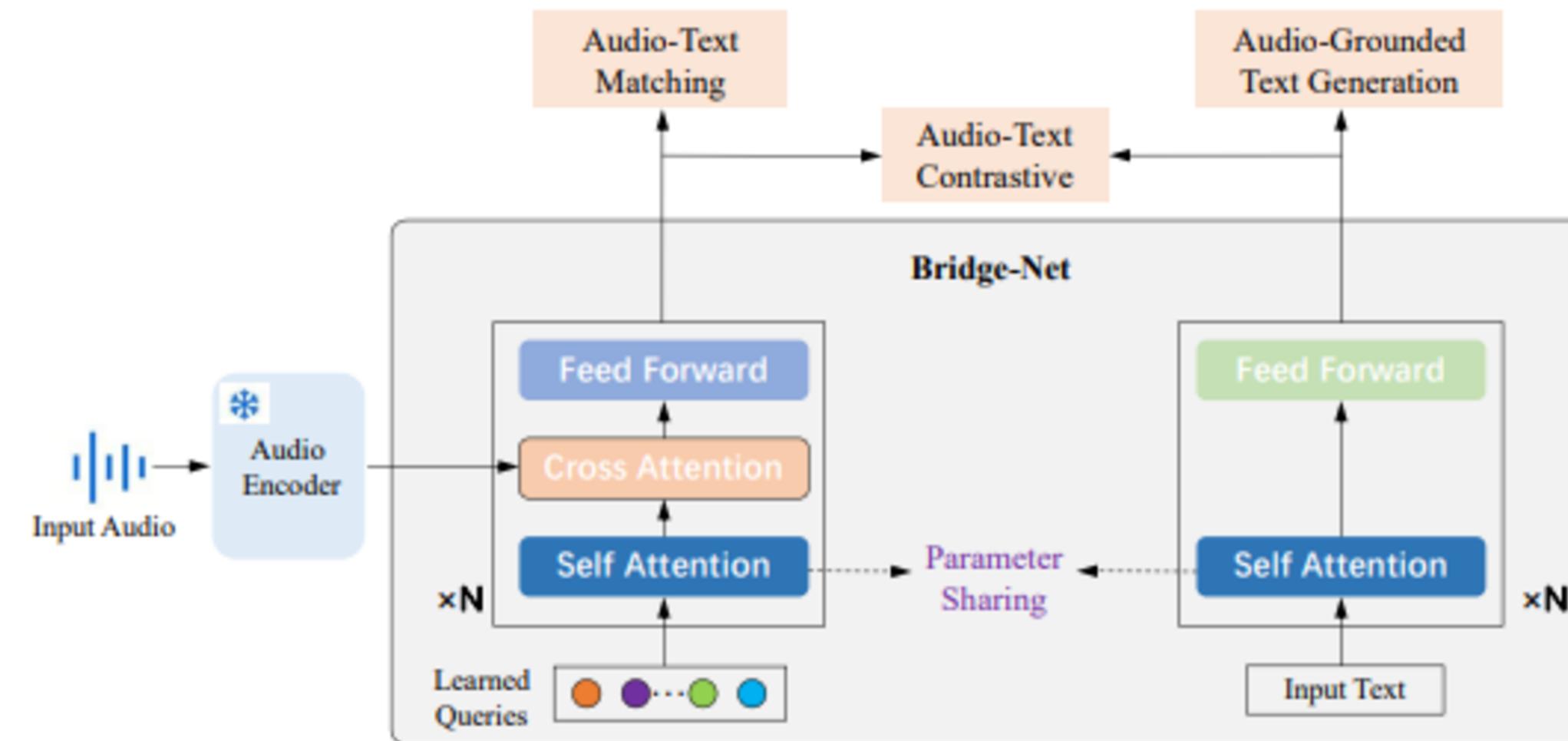


Figure 1: *Model architecture of our MINT's audio-language representation learning.*

# III. Downstream multimodal model

## 2. Speech LLM

### a) MINT (Bytedance -2024)

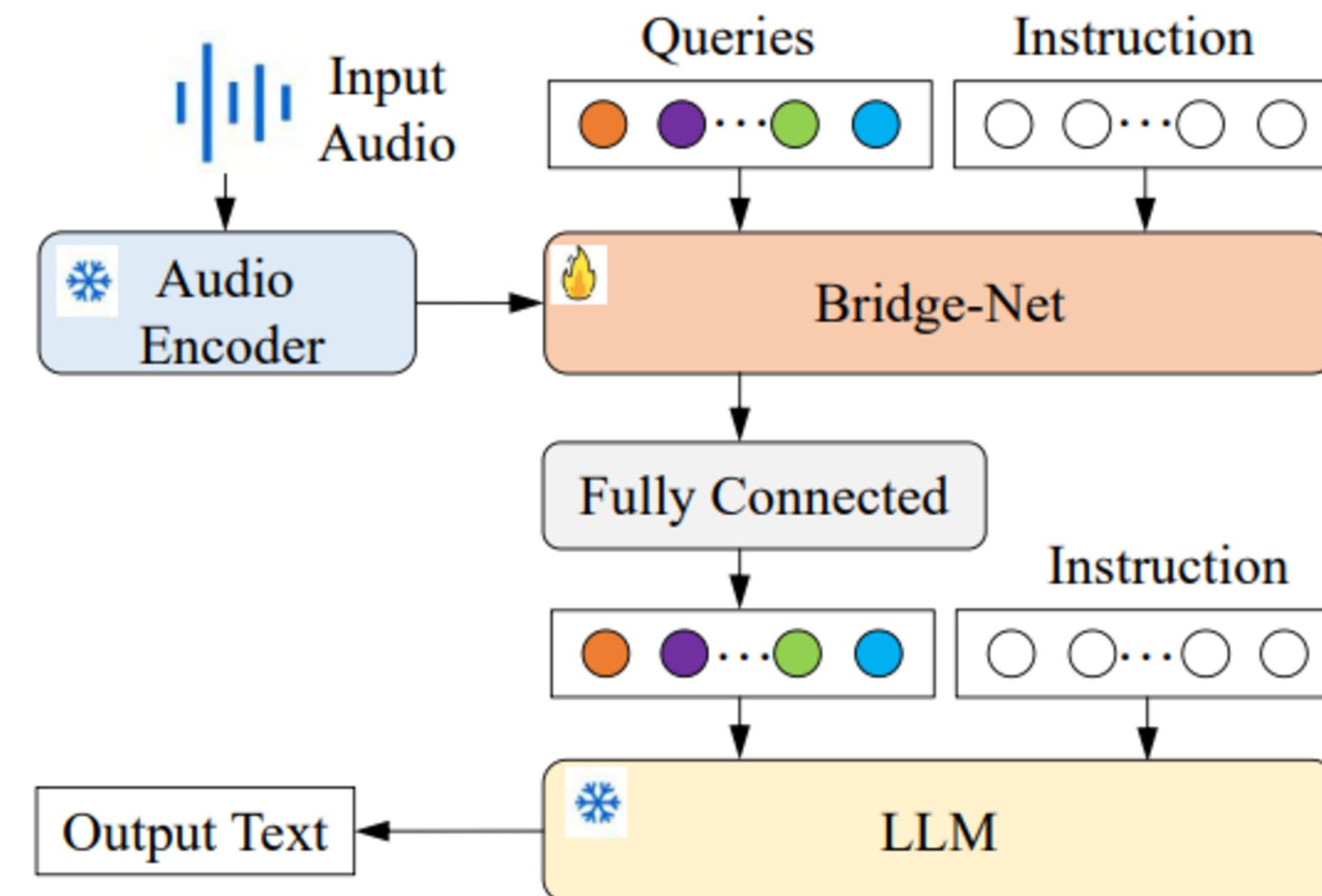
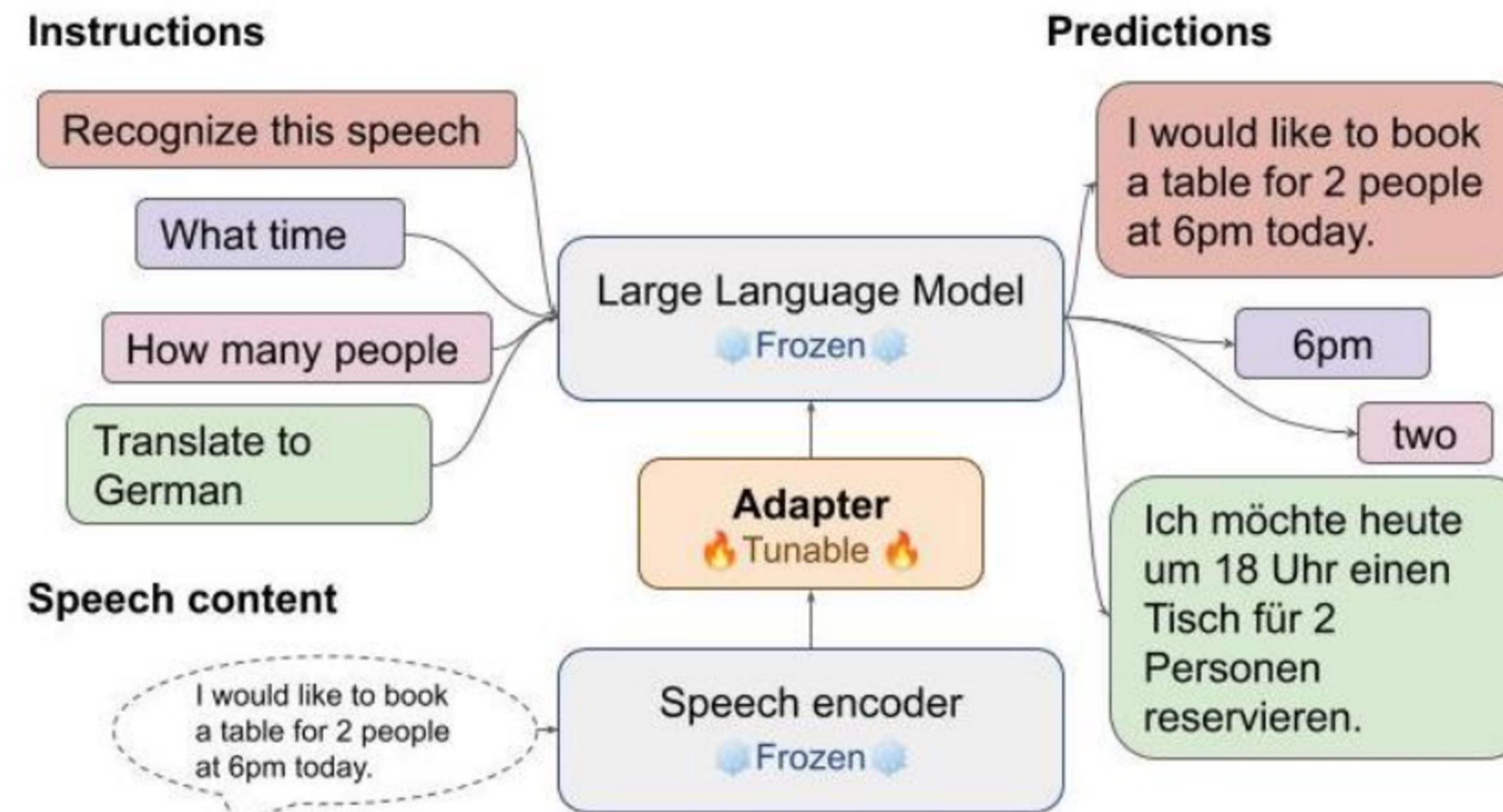


Figure 2: *MINT's instruction tuning process.*

# III. Downstream multimodal model

## 2. Speech LLM

b) SLM (Google Deepmind 2023)



# III. Downstream multimodal model

## 2. Speech LLM

b) SLM (Google Deepmind 2023)

