

# MCAT: Scaling Many-to-Many Speech-to-Text Translation with MLLMs to 70 Languages

Yexing Du, Kaiyuan Liu, Youcheng Pan, Bo Yang, Keqi Deng,  
Xie Chen, Yang Xiang, Ming Liu, Bin Qin, YaoWei Wang

**Abstract**—Multimodal Large Language Models (MLLMs) have achieved great success in Speech-to-Text Translation (S2TT) tasks. However, current research is constrained by two key challenges: language coverage and efficiency. Most of the popular S2TT datasets are substantially English-centric, which restricts the scaling-up of MLLMs’ many-to-many translation capabilities. Moreover, the inference speed of MLLMs degrades dramatically when the speech is converted into long sequences (e.g., 750 tokens). To address these limitations, we propose a Multilingual Cost-effective Accelerated Speech-to-Text Translator (MCAT) framework, which includes two innovations. First, a language scaling method that leverages curriculum learning and a data balancing strategy is introduced to extend the language coverage supported by MLLMs to 70 languages and achieve mutual translation among these languages. Second, an optimized speech adapter module is designed to reduce the length of the speech sequence to only 30 tokens. Extensive experiments were conducted on MLLMs of different scales (9B and 27B). The experimental results demonstrate that MCAT not only surpasses state-of-the-art end-to-end models on the FLEURS dataset across  $70 \times 69$  directions but also enhances batch inference efficiency. This is achieved with only  $\sim 100M$  trainable parameters and by using only 10 hours of S2TT data per language. Furthermore, we have released MCAT as open-source to promote the development of MLLMs for robust S2TT capabilities.<sup>12</sup>

**Index Terms**—Speech-to-Text Translation, Multimodal Large Language Models, Curriculum Learning.

## I. INTRODUCTION

Speech-to-Text Translation (S2TT) involves converting speech from a source language into text in a target language. Traditionally, S2TT tasks have relied on a cascaded system, where an Automatic Speech Recognition (ASR) model first transcribes the speech into text [2], followed by a Machine Translation (MT) model that translates the text into the target language [3]. Recently, MLLMs [4] have demonstrated ad-

Manuscript created November, 2025. (Corresponding authors: Ming Liu; Yang Xiang.)

Yexing Du, Kaiyuan Liu and Yaowei Wang are with Harbin Institute of Technology, Shenzhen, China, and also with Pengcheng Laboratory, Shenzhen, China (e-mail: yxdu@ir.hit.edu.cn; 1171000408@stu.hit.edu.cn; wangyaowei@hit.edu.cn).

Ming Liu and Bin Qin are with Harbin Institute of Technology, Harbin, China and also with Pengcheng Laboratory, Shenzhen, China (e-mail: mliu@ir.hit.edu.cn; qinb@ir.hit.edu.cn).

Youcheng Pan, Bo Yang and Yang Xiang are with Pengcheng Laboratory, Shenzhen, China (e-mail: panych@pcl.ac.cn; yangb05@pcl.ac.cn; xiangy@pcl.ac.cn).

Keqi Deng is with University of Cambridge, CB2 1TN Cambridge, U.K (e-mail: kd502@cam.ac.uk).

Xie Chen is with Shanghai Jiao Tong University, Shanghai, China (e-mail: chenxie95@sjtu.edu.cn).

<sup>1</sup>The code and models are released at <https://github.com/yxduir/m2m-70>.

<sup>2</sup>This manuscript is an extended version of [1].

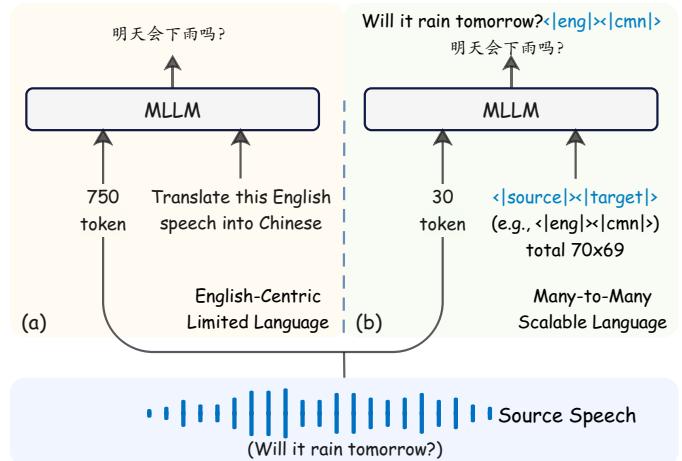


Fig. 1. Comparison of S2TT MLLMs. (a) compresses speech to 750 tokens, has limited language support, and directly generates translated text; (b) generates transcriptions and translations in a single end-to-end pass, compressing speech to 30 tokens, supporting 70 languages.  $<|eng|><|cmn|>$  indicates transcribing English and translating it into Chinese.

vantages in simplifying the model architecture and mitigating error propagation [5] in both ASR [6] and S2TT tasks [7].

However, existing MLLMs for S2TT are constrained by two challenges: **language coverage** and **efficiency**. First, MLLM training is usually data-driven, but the existing S2TT datasets [8] are predominantly English-centric. This leads to limited language coverage and weak many-to-many translation capabilities, as shown in Figure 1(a). Second, current MLLMs often employ an adapter structure similar to LLaVA [9], which uses an MLP directly to project features into the LLM, resulting in a very long input sequence (e.g., 750 tokens [4]), even for extremely short samples such as “Will it rain tomorrow?”, leading to limited inference efficiency.

To address these limitations, this research presents two key innovations. First, we introduce a **language scaling** strategy that includes a three-stage curriculum learning strategy (utilizing ASR data for pre-training, and minimal S2TT data to establish the connection between MT and S2TT), and a data balancing strategy to handle multilingual data imbalance. Finally, we extend the MLLM’s S2TT task support to mutual translation among 70 languages, as shown in Figure 1(b). Second, we design an efficient **speech adapter** structure, which utilizes a Q-Former [10] for feature extraction, pooling for compression, and an MLP for aligning the features to the LLM’s dimension. This design reduces the speech token input to just 30 tokens.

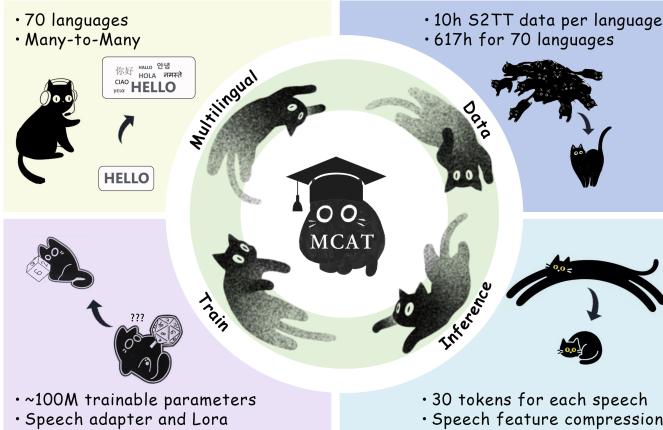


Fig. 2. **Key Features:** (a) Multilingual Support; (b) Low-Resource Requirement; (c) Lightweight Training; (d) High-Efficiency Inference.

Based on the above design, Multilingual Cost-effective Accelerated Speech-to-Text Translator (MCAT) models exhibit four key features shown in Figure 2: (1) Multilingual Support, offering many-to-many S2TT across **70** languages; (2) Low-Resource Requirement, needing only **10** hours of S2TT data per language; (3) Lightweight Training, achieved by utilizing the speech adapter and LoRA for efficient parameter training ( $\sim 100M$  trainable parameters); and (4) High-Efficiency Inference, enabled by compressing the input speech sequences to just **30** tokens to accelerate batch inference.

To evaluate the impact of our proposed methods, we trained two MLLM variants of different scales (**9B** and **27B**). Crucially, in low-resource settings, our models demonstrated powerful many-to-many S2TT capability on the FLEURS [11] dataset across 70 languages, outperforming existing state-of-the-art end-to-end models. Furthermore, we also validated the data scaling law on the CoVoST-2 [8] dataset. Finally, our strategies were validated by comprehensive ablation studies and comparisons of inference speed.

Our main contributions are summarized as follows:

- We introduce a **Language Scaling Strategy** (including curriculum learning and data balancing) to enable many-to-many S2TT support across **70** languages, with comprehensive evaluation and analysis conducted across all **4,830** directions.
- We propose an optimized **Speech Adapter** that achieves an extreme  **$25\times$**  input compression, reducing the speech sequence length to **30** tokens. Despite such extreme compression, our model still achieved state-of-the-art end-to-end S2TT performance on FLEURS dataset.
- We validate that our MCAT framework is highly **Data- and Parameter-Efficient**. Our models (**9B** and **27B**) achieve superior performance by fine-tuning only  $\sim 100M$  parameters and utilizing minimal S2TT data (< 10 hours per language) for language extension.

In this paper, we extend our earlier work at ACL 2025 [1]. Specifically, we introduce a language scaling strategy to scale up the multilingual support from 15 to **70 languages**. Furthermore, we refine the speech adapter architecture to reduce the number of speech tokens to just **30**.

## II. RELATED WORK

### A. Cascaded S2TT

The Cascaded S2TT approach typically employs a two-step pipeline: Automatic Speech Recognition (ASR) first transcribes the source spoken language into text, and subsequently Machine Translation (MT) translates the transcribed text into the target language. Specifically, established ASR models, such as Whisper [12], accurately convert speech into text. Similarly, MT models, for instance NLLB [13], achieve high translation accuracy and fluency by utilizing large multilingual datasets. However, a significant limitation of the cascaded approach is its susceptibility to error propagation.

### B. End-to-End S2TT

Distinct from the cascaded paradigm, End-to-End S2TT trains a unified model to directly map speech from the source language to text in the target language, thereby eliminating the intermediate transcription step [14]. Certain models, like Whisper [12], also support multilingual-to-English translation capabilities. Furthermore, models such as SeamlessM4T-V2-Large [15] represent strong encoder-decoder architectures for diverse multilingual speech-to-text tasks. These pioneering efforts often prioritize reducing latency and enhancing efficiency over traditional offline speech translation systems.

### C. Audio MLLMs

Recently, the rapid advancements in MLLMs [16] have substantially improved performance in speech recognition and translation tasks. Approaches like SpeechGPT [6] utilize prompting mechanisms to enhance speech recognition within large language models. SALMONN [17] specifically focuses on improving the auditory comprehension of both language and music. Qwen-Audio [4] advances audio recognition and translation by retraining speech encoders within a multi-task framework. More recently, Voxtral [18] and Qwen3-Omni [19] have further extended this progress by integrating enhanced multimodal understanding. In addition, LLM-SRT [1] introduces a curriculum learning strategy designed to strengthen cross-modal alignment and translation quality.

TABLE I  
S2TT LANGUAGE COVERAGE.

S2TT Models	Language	S2TT Data (h)
<b>Encoder-Decoder Models</b>		
Whisper-Large-V2 [12]	96 → eng	125,000
SeamlessM4T-V2-Large [15]	101 ↔ 96	351,000
<b>MLLMs</b>		
Qwen-Audio-7B [4]	6 ↔ 6	3,700
Voxtral-Small-24B [18]	8 ↔ 8	in-house
Qwen3-Omni-30B-A3B-Instruct [19]	19 ↔ 19	in-house
MCAT-Small-9B (ours)	28 ↔ 28	243.9
MCAT-Large-27B (ours)	70 ↔ 70	617.7

Language coverage follows reported S2TT scores in the paper.

### III. METHODOLOGY

#### A. Problem Formulation

In this section, we define the following tasks:

- **Automatic Speech Recognition (ASR):** Given the speech input  $x$  and the instruction text  $t$ , the goal is to produce the transcribed text  $Y_1$ .
- **Speech-guided Machine Translation (SMT):** Given the speech input  $x$ , its transcription  $Y_1$ , and the instruction text  $t$ , the goal is to produce the translated text  $Y_2$ .
- **Speech Recognition and Translation (SRT):** Given the speech input  $x$  and the instruction text  $t$ , the goal is to produce the transcription  $Y_1$  and the translation  $Y_2$ .

#### B. Model Architecture

As detailed in Table IV, the MCAT models are built upon an LLM. They adopts Whisper's encoder [12] as the speech encoder, followed by a Q-Former [10], Pooling, and MLP layer for the speech adapter. Notably, our design compresses **30 seconds of speech into 30 tokens** to boost MLLM inference efficiency.

1) **Speech Preprocessing:** The raw waveform  $x \in \mathbb{R}^{N \times T}$  ( $N$  being the batch size and  $T$  the temporal length) undergoes audio processing, including STFT and Mel Filterbanks, to convert the time-domain signal into a Mel-spectrogram  $M$ .

$$x \in \mathbb{R}^{N \times T} \xrightarrow[\text{STFT}]{\text{Mel Filterbanks}} M \in \mathbb{R}^{N \times C \times L}, \quad (1)$$

where the Whisper encoder requires the input to be a fixed length  $L$ , achieved by truncation or padding. The dimension  $C$  represents the key feature size of the Mel-spectrogram.

2) **Speech Encoder:** We leverage the frozen Whisper's encoder, which maps the padded Mel-spectrogram input  $X$  to a sequence of hidden representations  $H$ :

$$H = \text{Encoder}(M), \quad H \in \mathbb{R}^{N \times L' \times D_w} \quad (2)$$

where  $D_w$  is the encoder hidden dimension and  $L'$  is the encoder sequence length.

3) **Speech Adapter:** The speech adapter layer comprises a Q-Former, a pooling layer, and an MLP. The Q-Former is responsible for feature extraction, the pooling layer handles feature compression, and the MLP layer performs dimension alignment with the LLM's embedding space.

a) **Q-Former for Feature Extraction:** The Q-Former serves to extract a set of compact, fixed-length speech embeddings  $Z$  from the longer sequence  $H$ :

$$Z = \text{Q-Former}(H), \quad Z \in \mathbb{R}^{N \times K \times D_q} \quad (3)$$

Here,  $K$  is the fixed number of learned query tokens, and  $D_q$  is the Q-Former's hidden dimension.

b) **Pooling Layer for Feature Compression:** Temporal pooling is applied to reduce the sequence length by  $S \times$  downsampling (e.g., average pooling):

$$Z_p = \text{Pool}(Z), \quad Z_p \in \mathbb{R}^{N \times K / S \times D_q} \quad (4)$$

This operation compresses the 150 acoustic features to **30**.

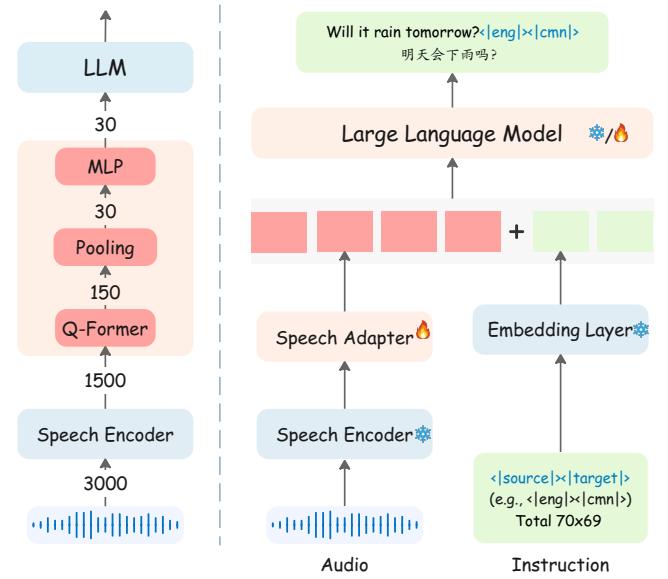


Fig. 3. **The Architecture of MCAT Model.** Our MLLM compresses the input audio into 30 tokens, supporting a total of 70 languages.

c) **MLP for Dimension Alignment:** The MLP maps the compressed features into the LLM's dimension  $D_{llm}$ :

$$Z_{mlp} = \text{MLP}(Z_p), \quad Z_{mlp} \in \mathbb{R}^{N \times K / S \times D_{llm}} \quad (5)$$

where  $Z_{mlp}$  represents the aligned speech feature embeddings, ready for concatenation.

4) **Text Embedding:** Given the instruction text  $t$ , the corresponding prompt embeddings are obtained as:

$$P = \text{Embedding}(t) \in \mathbb{R}^{N \times P_t \times D_{llm}}, \quad (6)$$

where  $P_t$  is the prompt token length.

5) **Multimodal Fusion and LLM Output:** To achieve multimodal integration, the modality-specific features  $Z_{mlp}$  are fused with the text embeddings  $P$  by concatenating them along the temporal dimension:

$$X = \text{Concat}(Z_{mlp}, P) \in \mathbb{R}^{N \times (K / S + P_t) \times D_{llm}} \quad (7)$$

The fused representation  $X$  is subsequently fed into the LLM, which autoregressively produces the text outputs  $Y$ .

TABLE II  
STAGES AND OUTPUT SHAPES

Input	Stage	Feature	Shape
Speech	Raw Speech	$x \in \mathbb{R}^{N \times T}$	$N \times T$
	Mel-spectrogram	$M \in \mathbb{R}^{N \times C \times L}$	$N \times 128 \times 3000$
	Speech Encoder	$H \in \mathbb{R}^{N \times L' \times D_w}$	$N \times 1500 \times 1280$
	Q-Former	$Z \in \mathbb{R}^{N \times K \times D_q}$	$N \times 150 \times 768$
	Pooling	$Z_p \in \mathbb{R}^{N \times K / S \times D_q}$	$N \times 30 \times 768$
	MLP	$Z_{mlp} \in \mathbb{R}^{N \times K / S \times D_{llm}}$	$N \times 30 \times D_{llm}$
Text	Text Embedding	$P \in \mathbb{R}^{N \times P_t \times D_{llm}}$	$N \times P_t \times D_{llm}$
LLM Input	Multimodal Fusion	$X \in \mathbb{R}^{N \times (K / S + P_t) \times D_{llm}}$	$N \times (30 + P_t) \times D_{llm}$

TABLE III  
INSTRUCTION DESIGN.

Task	Speech	Instruction Text	Prediction
ASR	✓	<  eng  >	Will it rain tomorrow?
	✓	<  deu  >	Regnet es morgen?
SMT	✓	Will it rain tomorrow?<  eng  ><  deu  >	Regnet es morgen?
	✓	Regnet es morgen?<  deu  ><  fra  >	Il va pleuvoir demain ?
SRT	✓	<  eng  ><  deu  >	Will it rain tomorrow?<  eng  ><  deu  >Regnet es morgen?
	✓	<  deu  ><  fra  >	Regnet es morgen?<  deu  ><  fra  >Il va pleuvoir demain ?

### C. Language Scaling Strategy

MCAT employs a comprehensive language scaling strategy to effectively train an MLLM for multilingual S2TT across 70 languages. This strategy involves a three-stage curriculum learning strategy to bridge the connection between MT and S2TT tasks and a data balancing strategy focusing on balanced ASR and S2TT data usage.

*1) Language Tags:* Minimalist instructions are designed to help the model distinguish between tasks while minimizing instruction token length, as shown in Table III. This design ensures that task-specific markers, such as <| eng |><| deu |>, appear in the generated answers, effectively segmenting transcription and translation content in the SRT task.

*2) Curriculum Learning Strategy:* We adopt a curriculum learning approach that incorporates three sequential training tasks: ASR, SMT, and SRT. This sequence is designed to utilize data-rich ASR as a bridge to develop fundamental capabilities before scaling to the more complex SMT and SRT tasks.

*a) ASR Pre-training:* In this initial stage, the model is pre-trained to develop ASR capabilities with a focus on multi-modal alignment. This step also involves expanding language support by training on all intended languages. The speech adapter is trained with as much data as possible to ensure efficient fine-tuning and establish a strong foundation.

*b) SMT Enhancement:* This stage enhances the model's cross-lingual abilities. Starting from the ASR checkpoint, the model takes both transcribed text and audio as input to generate translations based on the instruction. The purpose is to activate the LLM's inherent MT capabilities and establish the necessary connection between the MT and the S2TT tasks.

*c) SRT Activation:* The final stage activates the model's full SRT capabilities. Training continues from the SMT checkpoint, with the model receiving only audio input and a task-specific instruction, outputting both the transcription and translation of the speech. This step extends the MT capabilities of LLMs to the S2TT task and finalizes the model.

*3) Data Balancing Strategy:* A core challenge when training our 70-language model is mitigating disparity in performance caused by inherent data imbalance. We employ a strategy that scales the language set from high-resource to low-resource languages, followed by a final balancing step.

*a) ASR Language Expansion:* Training begins with English and Chinese ASR data for foundational capability. The

### Algorithm 1: Language Scaling Strategy

```

Input: Initial Model  $\mathcal{M}_0$ 
Output: Final Model  $\mathcal{M}_{\text{final}}$ 
 $\mathcal{L}_{\text{full}} \leftarrow \text{All Languages};$ 
 $\mathcal{M}_{\text{ASR}} \leftarrow \mathcal{M}_0;$ 
Phase 1: ASR Pre-training // Activate ASR
  Capabilities
begin
   $\mathcal{L}_{\text{set}} \leftarrow \langle \mathcal{L}_2, \mathcal{L}_{28}, \mathcal{L}_{44}, \mathcal{L}_{\text{full}} \rangle;$ 
  // Language Sets
  for  $\mathcal{L}_{\text{subset}} \text{ in } \mathcal{L}_{\text{set}}$  do
     $\mathcal{D}_{\text{subset}} \leftarrow \text{GetASRData}(\mathcal{L}_{\text{subset}})$ 
     $\mathcal{M}_{\text{ASR}} \leftarrow \text{FineTune}(\mathcal{M}_{\text{ASR}}, \mathcal{D}_{\text{subset}})$  // ASR
      (Audio → Transcription)
  end

Phase 2: Balanced ASR Fine-Tuning begin
   $\mathcal{D}_{\text{BalancedASR}} \leftarrow \text{GetASRData}(\mathcal{L}_{\text{full}}, \text{Max} = 10000)$ 
  // Limit to 10,000 samples per language
   $\mathcal{M}_{\text{ASR}} \leftarrow \text{FineTune}(\mathcal{M}_{\text{ASR}}, \mathcal{D}_{\text{BalancedASR}})$ 
Phase 3: SMT and SRT Full-Scale Training
  // Activate S2TT Capabilities
begin
   $\mathcal{D}_{\text{SMT}} \leftarrow \text{GetSMTData}(\mathcal{L}_{\text{full}})$ 
   $\mathcal{M}_{\text{SMT}} \leftarrow \text{FineTune}(\mathcal{M}_{\text{ASR}}, \mathcal{D}_{\text{SMT}})$  // SMT (Audio + Transcription → Translation)
   $\mathcal{D}_{\text{SRT}} \leftarrow \text{GetSRTData}(\mathcal{L}_{\text{full}})$ 
   $\mathcal{M}_{\text{SRT}} \leftarrow \text{FineTune}(\mathcal{M}_{\text{SMT}}, \mathcal{D}_{\text{SRT}})$  // SRT (Audio → Transcription + Translation)
Phase 4: Balanced SRT Fine-Tuning begin
   $\mathcal{D}_{\text{BalancedSRT}} \leftarrow \text{GetSRTData}(\mathcal{L}_{\text{full}}, \text{Max} = 100)$  // Limit to 100 samples per direction
   $\mathcal{M}_{\text{final}} \leftarrow \text{FineTune}(\mathcal{M}_{\text{SRT}}, \mathcal{D}_{\text{BalancedSRT}})$ 

```

language set is then progressively expanded in stages: first to 28 languages, then to 44, and finally to the full 70 languages.

*b) Balanced ASR Fine-Tuning:* In this stage, we reduced the ASR data for all languages to a maximum of 10,000 samples per language, and then continued ASR training based on the previous checkpoint.

*c) SMT and SRT Full-Scale Training:* We continue training from the ASR checkpoint using data from all 70 languages to enhance S2TT capabilities. The model is first fine-tuned on the SMT task, then on the SRT task.

*d) Balanced SRT Fine-Tuning:* In this stage, we reduced the SRT data for all language directions to a maximum of 100 samples per direction, and then continued SRT training based on the previous checkpoint.

## IV. EXPERIMENTS SETTING

### A. Datasets

In our experiments, we use the *CommonVoice*<sup>3</sup> [20] and *FLEURS*<sup>4</sup> [11] datasets for the **ASR** task training, and the *FLEURS* dataset for the **SMT** and **SRT** task training. We perform comparative and ablation studies on the *FLEURS* and *CoVoST-2*<sup>5</sup> [8] datasets. Detailed information for datasets is provided in Table XII.

### B. Model Architecture

As shown in Table IV, the MLLM consists of an LLM (GemmaX2-9B [21] or Gemma3-27b-it [22]), a frozen speech encoder (Whisper-large-v3), and a trainable adapter layer comprising a Q-Former, Pooling layer and MLP layer. For Q-Former, we use 150 queries, each with a dimension of 768. Training can be minimized by freezing the LLM, or LoRA [23] can be applied for training.

### C. Training Details

We used BF16 precision with Distributed Data Parallel (DDP), a learning rate of  $5 \times 10^{-5}$ , 1000 warmup steps, and the AdamW optimizer. The models were trained on 8 A100 GPUs. The 9B model can be trained in 3 days, while the 27B model can be trained in 7 days.

### D. Compared Methods

We compare both cascade systems and end-to-end S2TT models, such as SeamlessM4T [24] which supports S2TT for nearly 100 languages, and Qwen-Omni series [19], the open-source MLLM that centers on English and Chinese, extending its capabilities to diverse audio modalities.

### E. Language Support

As shown in Table V, our MCAT-Small model supports 28 languages across 9 language families, while the MCAT-Large model supports 70 languages across 12 language families. Instructions for language support can be found in the appendix.

TABLE IV  
MLLM TRAINING SETTINGS.

Modules	MCAT-Small	MCAT-Large	Train Stage	Details
Speech Encoder	~635M	~635M	-	Whisper's encoder
Speech Adapter	~80.6M	~85.2M	All	Q-Former / Pooling / MLP
LLM	~9.2B	~27.4B	-	GemmaX2-9B or Gemma3-27B
LLM Lora	~8.9M	~18.7M	MCAT	LoRA (r=16, alpha=32)
Total Trainable	~89.5M	~103.9M		
Total	~10B	~28B		

The blue color indicates the trainable parameters.

TABLE V  
LANGUAGE SUPPORT.

ISO Code	Language	Family	MCAT-Small	MCAT-Large	S2TT Data (h)
afr	Afrikaans	Indo-European		✓	3.6
amh	Amharic	Afro-Asiatic		✓	11.1
ara	Arabic	Afro-Asiatic	✓	✓	6.0
asm	Assamese	Indo-European		✓	10.7
azj	Azerbaijani	Turkic		✓	9.3
bel	Belarusian	Indo-European		✓	9.5
ben	Bengali	Indo-European	✓	✓	10.7
bos	Bosnian	Indo-European		✓	10.0
bul	Bulgarian	Indo-European		✓	9.5
cat	Catalan	Indo-European		✓	7.4
ces	Czech	Indo-European	✓	✓	8.4
cmm	Chinese	Sino-Tibetan	✓	✓	9.7
cym	Welsh	Indo-European		✓	12.2
dan	Danish	Indo-European	✓	✓	7.5
deu	German	Indo-European		✓	9.0
ell	Greek	Indo-European	✓	✓	10.0
eng	English	Indo-European	✓	✓	7.5
est	Estonian	Uralic		✓	7.3
fas	Persian	Indo-European	✓	✓	12.1
fin	Finnish	Uralic		✓	8.8
fra	French	Indo-European	✓	✓	10.3
glg	Galician	Indo-European		✓	6.7
guj	Gujarati	Indo-European		✓	9.0
heb	Hebrew	Afro-Asiatic	✓	✓	9.5
hin	Hindi	Indo-European	✓	✓	6.7
hrv	Croatian	Indo-European		✓	11.8
hun	Hungarian	Uralic		✓	9.3
hye	Armenian	Indo-European		✓	10.4
ind	Indonesian	Austronesian	✓	✓	9.1
isl	Icelandic	Indo-European		✓	2.8
ita	Italian	Indo-European	✓	✓	9.0
jav	Javanese	Austronesian		✓	11.2
jpn	Japanese	Japonic	✓	✓	7.4
kan	Kannada	Dravidian		✓	8.3
kat	Georgian	Kartvelian		✓	5.1
kaz	Kazakh	Turkic		✓	11.8
khm	Khmer	Austroasiatic	✓	✓	7.1
kir	Kyrgyz	Turkic		✓	9.3
kor	Korean	Koreanic	✓	✓	7.9
lao	Lao	Kra-Dai	✓	✓	7.3
lav	Latvian	Indo-European		✓	6.5
lit	Lithuanian	Indo-European		✓	9.8
mal	Malayalam	Dravidian		✓	10.1
mkd	Macedonian	Indo-European		✓	6.8
msa	Malay	Austronesian	✓	✓	9.5
mya	Burmese	Sino-Tibetan		✓	12.1
nld	Dutch	Indo-European	✓	✓	7.7
nob	Norwegian	Indo-European		✓	10.9
npi	Nepali	Indo-European		✓	11.3
pan	Punjabi	Indo-European		✓	6.4
pol	Polish	Indo-European	✓	✓	9.2
por	Portuguese	Indo-European	✓	✓	10.2
ron	Romanian	Indo-European		✓	10.1
rus	Russian	Indo-European	✓	✓	8.1
slk	Slovak	Indo-European		✓	5.9
slv	Slovenian	Indo-European		✓	7.8
spa	Spanish	Indo-European	✓	✓	8.8
srp	Serbian	Indo-European		✓	10.7
swe	Swedish	Indo-European		✓	8.4
swh	Swahili	Niger-Congo		✓	13.5
tam	Tamil	Dravidian		✓	8.7
tel	Telugu	Dravidian		✓	7.9
tgl	Tagalog	Austronesian	✓	✓	7.7
tha	Thai	Kra-Dai	✓	✓	8.5
tur	Turkish	Turkic	✓	✓	8.3
ukr	Ukrainian	Indo-European		✓	9.0
urd	Urdu	Indo-European	✓	✓	7.0
uzb	Uzbek	Turkic		✓	10.1
vie	Vietnamese	Austroasiatic	✓	✓	9.1
yue	Cantonese	Sino-Tibetan		✓	7.3
Total				28	70
					617.7

### F. Evaluation Metrics

We use COMET<sup>6</sup> [25] and spBLEU<sup>7</sup> [26] as evaluation metrics. The spBLEU utilizes the FLORES-200 tokenizer. Instructions for evaluation metrics can be found in the appendix.

<sup>3</sup><https://datacollective.mozilla.org/datasets?q=common+voice>

<sup>4</sup><https://huggingface.co/datasets/google/fleurs>

<sup>5</sup><https://github.com/facebookresearch/covost>

<sup>6</sup><https://unbabel.github.io/COMET/>

<sup>7</sup><https://github.com/mjpost/sacrebleu>

TABLE VI  
COMET RESULTS ON  $9 \times 27$  AND  $9 \times 69$  DIRECTIONS ON THE FLEURS DATASET. SPBLEU RESULTS ARE SHOWN IN TABLE XIII.

Systems ( $X \rightarrow 27$ )	ara	cmn	eng	ind	jpn	kor	tha	tur	vie	Avg.
<b>Cascaded ASR+MT Models</b>										
Whisper-Large-V3 + NLLB-200-3.3B [13]	78.1	79.8	83.4	81.6	79.9	81.2	78.2	82.3	78.3	80.3
Whisper-Large-V3 + LLaMAX3-8B-Alpaca [27]	76.0	78.6	81.3	79.4	78.4	79.3	77.0	79.2	77.1	78.5
<b>End-to-end S2TT Models</b>										
SeamlessM4T-V2-Large [15]	69.4	72.6	84.3	71.2	69.1	73.5	68.6	71.2	71.9	72.4
MCAT-Small-9B (ours)	77.8	79.8	85.9	82.3	79.7	81.9	78.6	82.1	79.3	80.8
MCAT-Large-27B (ours)	78.7	80.3	86.3	83.2	79.8	82.4	78.0	83.2	79.5	81.3
Systems ( $X \rightarrow 69$ )	ara	cmn	eng	ind	jpn	kor	tha	tur	vie	Avg.
<b>Cascaded ASR+MT Models</b>										
Whisper-Large-V3 + NLLB-200-3.3B [13]	78.3	79.8	83.9	81.9	79.8	81.4	78.2	82.4	78.6	80.5
Whisper-Large-V3 + LLaMAX3-8B-Alpaca [27]	75.3	77.8	80.8	78.8	77.5	78.4	75.6	78.2	76.3	77.6
<b>End-to-end S2TT Models</b>										
SeamlessM4T-V2-Large [15]	70.2	73.9	85.2	71.7	69.4	74.0	69.1	72.2	72.9	73.2
MCAT-Large-27B (ours)	79.0	80.6	86.5	83.5	80.0	82.6	78.2	83.2	79.9	81.5

Underlined denotes previous state-of-the-art models, while highlighted entries surpass the previous models.

## V. EXPERIMENTS

### A. Overall Results

As shown in Tables VI, we evaluate the many-to-many S2TT performance across 70 languages on the FLEURS dataset. Table VII and Figure 4 show the performances for  $\text{eng} \rightarrow 27$  and  $\text{eng} \rightarrow 69$  directions with end-to-end models, respectively. Figure VIII shows the performances for 70 languages for MCAT-Large and SeamlessM4T-V2-Large. Table XI provides a comparison of inference speed, and Table IX presents an ablation study of the curriculum learning strategy.

### B. Many-to-Many S2TT on FLEURS

Table VI summarizes the COMET scores for S2TT on the FLEURS dataset. Our proposed models consistently demonstrate superior performance over prior end-to-end S2TT models. For the  $9 \times 27$  directions, MCAT-Large achieves the highest average COMET score of 81.3, significantly outperforming SeamlessM4T-V2-Large (72.4) by over 8 points. This superiority is maintained in the more challenging  $9 \times 69$  settings, where MCAT-Large secures the top average score of 81.5 (vs. 73.2), confirming the robustness and high-quality output of the SRT architecture across a wide range of language pairs.

1) **Language Direction:** Our models are configured in two variants: MCAT-Small-9B (supporting 28 languages) and MCAT-Large-27B (supporting 70 languages). To rigorously evaluate the translation performance, we identified **9 language families** commonly supported by both model variants. Then, we selected the language with the largest speaker population from each family. This methodology yielded two evaluation sets based on translation directions:  **$9 \times 27$  set** and  **$9 \times 69$  set** for the MCAT-Small and MCAT-Large models, respectively.

2) **English-centric vs. Balanced Optimization:** As shown in Table VI, SeamlessM4T-V2-Large exhibits strong performance in the English direction (e.g., 85.2 for  $9 \times 69$ ), suggesting an English-centric design that leads to a significant degradation of performance for other language pairs (average 73.2 for  $9 \times 69$ ). In contrast, MCAT-Large model demonstrates a much more balanced optimization across all nine representative source languages. While our English scores are highly competitive (e.g., 86.5 for  $9 \times 69$ ), our performance on non-English target languages is substantially higher across the board, resulting in a significantly better overall average (average 81.5 for  $9 \times 69$ ).

3) **Cascaded Systems vs. End-to-End Models:** As shown in Table VI, previous end-to-end models mainly showed a significant performance advantage only in the English direction, benefiting from end-to-end training with abundant S2TT data aligned with English. However, their performance in non-English directions was often inferior to that of cascaded systems. Our method successfully demonstrates the full performance advantage of MLLMs, thereby setting a new comprehensive state-of-the-art performance across all nine representative source language directions.

4) **Scaling Law of Language Coverage:** Typically, as the number of supported languages increases, MLLMs suffer from severe knowledge conflict and catastrophic forgetting. As shown in Table VI, MCAT-Large exhibits remarkably consistent performance across significant language expansion: achieving an average score of 81.3 across 27 languages and 81.5 across 69 languages. This near-perfect consistency strongly suggests that our data balancing and training strategy effectively mitigates language performance degradation under scale, successfully balancing performance while achieving strong language scalability.

TABLE VII  
COMET RESULTS ON ENG → 27 DIRECTIONS ON THE FLEURS DATASET.

End-to-end Models (eng → X)	ara	ben	ces	cnn	deu	fas	fra	heb	hin	ind	ita	jpn	khm	kor
SeamlessM4T-V2-Large [15]	84.5	<u>84.6</u>	88.0	79.7	84.8	84.6	85.3	<u>84.5</u>	<u>78.1</u>	89.0	85.1	84.7	<u>79.9</u>	85.1
Qwen2.5-Omni-7B [28]	84.5	63.0	81.9	86.5	85.0	73.4	84.6	66.1	58.7	86.2	84.2	88.3	37.8	84.7
Qwen3-Omni-30B-A3B-Instruct [19]	<u>86.6</u>	83.3	<u>89.0</u>	<u>88.3</u>	<u>86.8</u>	<u>85.5</u>	<u>87.3</u>	73.9	77.7	<u>90.9</u>	<u>87.3</u>	<u>90.8</u>	77.6	<u>89.7</u>
MCAT-Small-9B (ours)	85.4	<u>85.0</u>	<u>89.4</u>	87.0	85.7	<u>86.4</u>	85.6	86.6	<u>78.1</u>	89.2	86.0	89.8	<u>81.4</u>	88.3
MCAT-Large-27B (ours)	86.1	<u>85.2</u>	<u>90.1</u>	87.2	86.5	<u>86.9</u>	86.3	<u>87.2</u>	<u>78.9</u>	90.2	86.8	90.4	79.7	88.5
End-to-end Models (eng → X)	lao	msa	mya	nld	pol	por	rus	spa	tgl	tha	tur	urd	vie	Avg.
SeamlessM4T-V2-Large [15]	<u>81.4</u>	86.6	<u>85.7</u>	85.1	85.7	86.6	86.3	83.2	<u>82.1</u>	83.3	86.7	<u>79.4</u>	85.6	84.3
Qwen2.5-Omni-7B [28]	38.7	83.5	41.1	82.1	81.5	86.7	85.2	83.4	55.6	82.1	79.2	51.6	76.4	73.8
Qwen3-Omni-30B-A3B-Instruct [19]	80.2	<u>88.3</u>	71.9	<u>86.2</u>	<u>87.3</u>	<u>88.4</u>	<u>88.9</u>	<u>85.4</u>	80.1	<u>88.8</u>	<u>88.1</u>	78.3	<u>88.5</u>	<u>85.0</u>
MCAT-Small-9B (ours)	82.6	87.1	<u>87.0</u>	86.2	<u>87.6</u>	87.1	87.7	84.6	<u>82.7</u>	87.1	87.6	<u>80.8</u>	87.1	85.9
MCAT-Large-27B (ours)	<u>83.1</u>	87.5	85.0	<u>86.6</u>	<u>88.1</u>	87.8	88.8	85.2	<u>83.3</u>	87.7	<u>88.2</u>	<u>80.9</u>	87.8	<u>86.3</u>

Underlined denotes previous state-of-the-art models, while highlighted entries surpass the previous models.

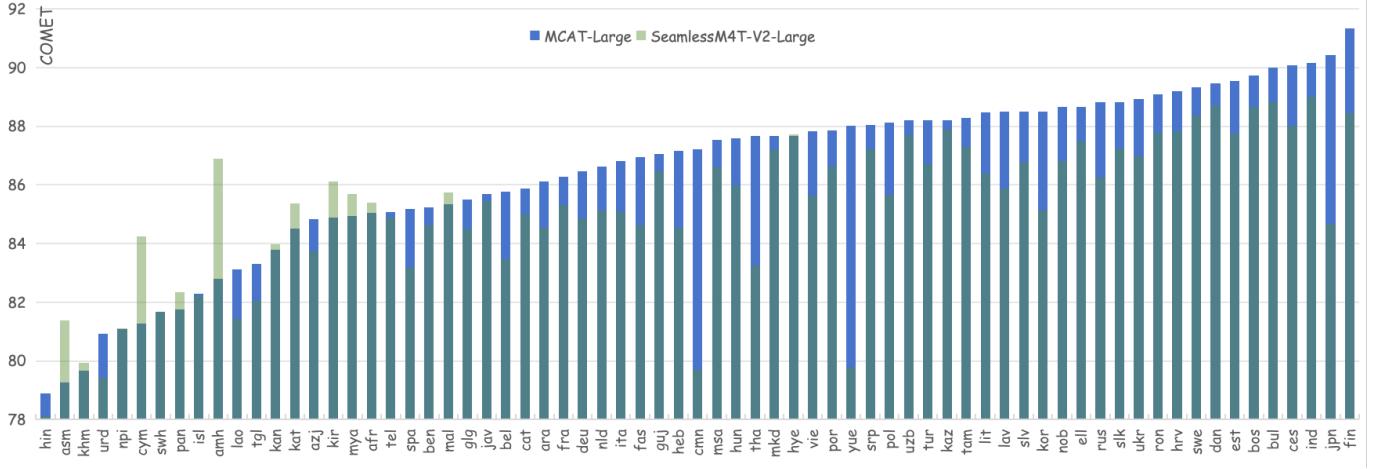


Fig. 4. COMET Scores for the English→69 Translation Directions on the FLEURS Dataset. The blue bars denote stronger translation performance for the MCAT-Large model in a total of 55 directions.

### C. Eng→X S2TT on FLEURS

Table IV presents a comprehensive comparison of the performance of various end-to-end translation models across 27 target language directions originating from English, evaluated using the COMET metric. Our proposed models, MCAT-Small and MCAT-Large, show competitive and often superior results compared to established models like SeamlessM4T-V2-Large and Qwen2.5-Omni-7B. Specifically, **MCAT-Large** achieves the highest overall average COMET score of **86.3**, surpassing the best prior model, Qwen3-Omni-30B-A3B-Instruct.

1) **Comparison on Eng→27 Language Directions:** As shown in Table VII, we compared the performance of end-to-end models on English. It can be observed that the models in the Qwen-Omni series show strong performance on high-resource languages such as **cnn** and **fra**, but exhibit noticeable deficiencies in low-resource languages like **khm** and **mya**. In contrast, our model achieves competitive performance on high-resource languages while demonstrating powerful performance on low-resource languages.

2) **Comparison on Eng → 69 Directions:** As Figure 4 illustrates, MCAT-Large shows a consistent performance advantage over the **SeamlessM4T-V2-Large** baseline, particularly in mid-to-high-resource settings (e.g., **tgl**, **cnn**). Quantitatively, MCAT-Large achieved superior results in 55 out of 69 tested directions, confirming a clear overall edge. However, its relatively weaker performance on low-resource languages (e.g., **amh**, **cym**) is primarily constrained by the intrinsic capabilities of the underlying LLM component within the MLLM architecture.

3) **MCAT-Small-9B vs. Qwen2.5-Omni-7B:** As shown in Figure 4, **MCAT-Small** consistently surpasses **Qwen2.5-Omni-7B** across all 27 translation directions, achieving a substantial average COMET gain of **11.2** points (from 73.8 to 85.0). Notably, its overall performance is comparable to that of Qwen3-Omni-30B-A3B-Instruct, underscoring the efficiency of our architecture. These results demonstrate that **MCAT-Small** provides competitive translation quality while maintaining lower computational and resource requirements.

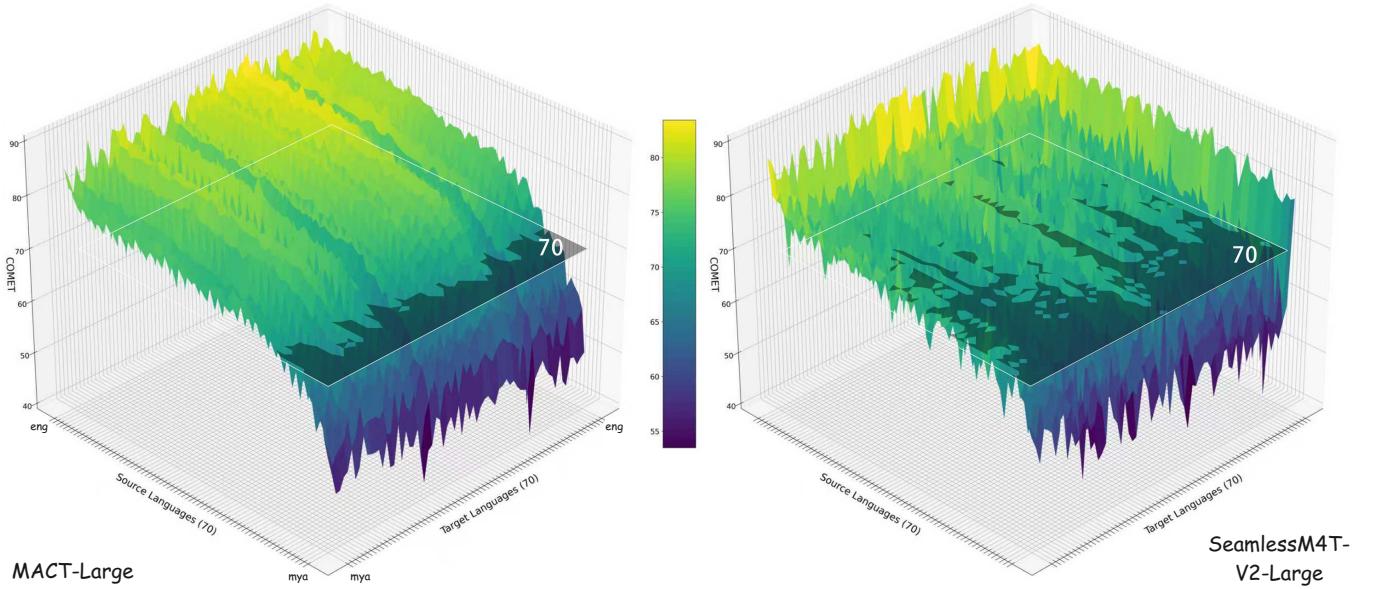


Fig. 5. COMET Scores Across  $70 \times 70$  Translation Directions. For cases like eng → eng, no score is calculated, and smoothing was applied in the figure.

TABLE VIII  
COMET SCORES STATISTICS ON THE FLEURS DATASET.

Models	$x \geq 90$	$90 > x \geq 80$	$80 > x \geq 70$	$x < 70$	Total
MCAT-Small-9B	0	399	265	92	$28 \times 27$
MCAT-Large-27B	6	2197	1834	793	$70 \times 69$
SeamlessM4T-V2-Large	0	215	2719	1896	$70 \times 69$

#### D. COMET Score Across 70 Languages

1) **Comparision on 70 Languages:** As shown in Figure 5, the surface is predominantly colored yellow and light green, corresponding to COMET scores well above 70. This signifies that the model provides usable translations across the vast majority of the potential language pairs. Furthermore, Figure 6 shows the translation performance of S2TT for 70 languages, ordered from smallest to largest average performance.

2) **Multilingual Consistency:** As shown in Figure 5, the MCAT model demonstrates a strong degree of multilingual consistency across translation directions. Specifically, for any given source language, the COMET scores when translating into the wide range of target languages are observed to be relatively uniform and cluster within a tight range. This consistency is a critical indicator of the model’s design success. It strongly suggests that the model is successfully employing shared knowledge components and parameter sharing across its multilingual capacity.

3) **Quantitative Confirmation of Robustness:** Table VIII provides a quantitative distribution of the COMET scores into specific score bins for two models. For the MCAT-Large-27B model, the majority of the  $70 \times 69$  directions (4,830 pairs) fall into the high-score brackets. Specifically, 4,037 pairs (combining the  $70 - 80$ ,  $80 - 90$ , and  $> 90$  bins) achieve a COMET score exceeding 70.

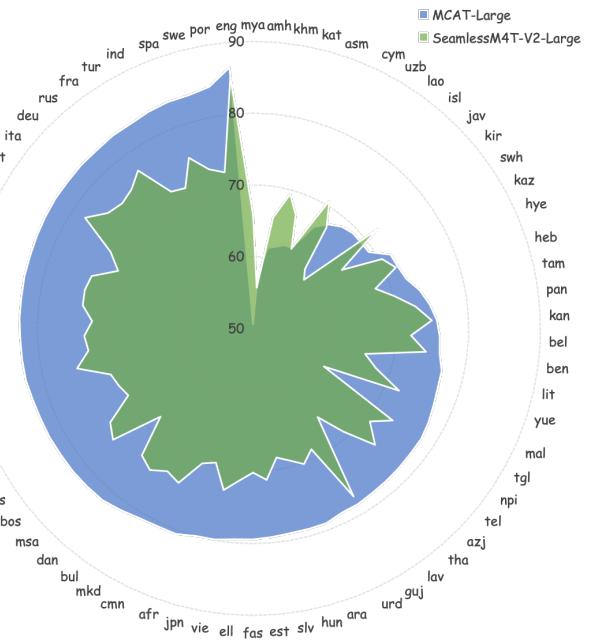


Fig. 6. Average Performance Across 70 Languages.

4) **Asymmetry in Low-Resource Language:** As shown in Figure 5, for languages such as **maya** (Burmese), **amh** (Amharic), and **khm** (Khmer), the COMET scores are very **high** when these languages serve as the **target language**. However, their scores are extremely **low** when they act as the **source language**, as shown in Figure 6. This suggests that the MLLM possesses sufficient capability to understand and generate text in these languages; however, the scarcity of speech recognition data prevents accurate speech decoding, leading to low overall scores. This finding implicitly suggests a critical need for more ASR data for these specific languages.

TABLE IX  
ABLATION STUDIES ON THE FLEURS DATASET.

spBLEU eng → X	eng S2TT Data (h)	ara	cmn	ind	jpn	khm	kor	lao	mya	tha	tur	vie	Avg.
MCAT-Small-9B	7.5	35.0	34.7	39.3	28.6	21.6	24.5	29.2	20.9	37.7	32.0	36.8	31.0
		14.8	15.7	20.6	10.3	12.2	8.3	13.9	9.1	19.2	12.7	18.2	14.1
		0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0
		34.5	34.5	38.9	27.8	21.6	24.4	28.9	20.4	37.5	32.2	37.0	30.7

### E. Ablation Study

The baseline model, MCAT-Small-9B, achieves robust performance in S2TT (eng→X), reporting an average spBLEU score of 31.0 across all eleven target languages on the FLEURS dataset. This strong result is underpinned by a foundational training setup that incorporates **7.5 hours** of English S2TT data. To confirm the impact of our strategy, we conduct ablation studies:

1) **Curriculum Learning Strategy:** The ablation study in Table IX conclusively demonstrates the necessity and superior efficacy of our multi-stage curriculum learning strategy. Eliminating this crucial component (w/o SMT+SRT) and reverting to a simple, direct instruction tuning regimen—similar to architectures like Qwen2-Audio [7]—results in a catastrophic performance drop: the average spBLEU score plummets from a robust 31.0 to a severely limited 14.1. This outcome provides definitive evidence that direct instruction tuning is wholly insufficient, particularly in an extremely low-data scenario, underscoring the curriculum’s role as an essential scaffold for gradual and robust knowledge acquisition.

2) **ASR Pretrain:** The ablation study in the row (Row: **w/o ASR**) demonstrates the critical importance of the ASR data component. When ASR data is removed, the model’s performance completely collapses, resulting in an average spBLEU score of **0.0**. This catastrophic failure strongly suggests that the ASR data is indispensable for the model to successfully learn the basic speech representation and robust audio encoding necessary for the downstream translation task, highlighting its role as the foundation of the curriculum learning approach. It serves as the pillar that establishes the core audio comprehension capability upon which the more complex task.

3) **Train Adapter Only vs. Fine-tune LLM:** We investigate the impact of fine-tuning the LLM component using LoRA. The baseline result from **w/o LLM Lora** shows that simply training the speech encoder (e.g., Q-Former) and the projector layer can achieve a surprisingly strong baseline performance. This indicates that the pre-trained LLM already possesses powerful cross-lingual reasoning and generation capabilities, requiring minimal adaptation to connect with the new speech features. However, the final fine-tuning step, **w/ LLM Lora**, provides a final marginal yet consistent performance gain across all evaluated language pairs. This confirms that while the primary knowledge transfer is handled by the adapter, modest, parameter-efficient tuning of the LLM’s weights is still beneficial for optimal integration and maximum translation accuracy.

TABLE X  
SCALING LAW OF DATA.

COMET eng→X	CoVoST-2							Avg.
	ara	cmn	deu	fas	ind	jpn	tur	
<b>with FLEURS eng data: 7.5 h</b>								
MCAT-Small-9B	79.8	82.0	79.7	80.3	83.7	85.1	81.3	81.7
<b>with CoVoST-2 eng data: 429.6 h</b>								
MCAT-Small-9B-V2	83.8	86.0	84.3	83.8	88.2	87.9	85.4	85.6

4) **Scaling Law of Data:** Table X validates the **Data Scaling Law**: increasing the English training data for the MCAT-Small-9B model from the low-data FLEURS regime (**7.5 h**) to the larger CoVoST-2 dataset (**429.6 h**) results in a dramatic performance surge. Specifically, this  $\sim 57\times$  increase in data volume boosts the average COMET score from 81.7 to **85.6**, confirming that even with a fixed architecture, performance is strongly bounded by the scale of the training data.

5) **Inference Speed:** As shown in Table XI, our **MCAT models** demonstrate superior computational efficiency compared to the Qwen2.5-Omni-7B under the same BF16 A100 GPU setup, despite the models utilizing Whisper encoders with vastly different token lengths (750 tokens vs. **30 tokens for a speech sample**). Specifically, the MCAT-Small-9B model achieves a remarkable inference time of only 76 seconds when utilizing 4 GPUs (batch 50), which is a **3.3×** speedup over the 253 seconds required by Qwen2.5-Omni-7B (4 GPUs, VLLM dynamic batch). Even the larger MCAT-Large-27B model remains significantly faster, completing the task in 169 seconds, validating the highly optimized architecture.

TABLE XI  
INFERENCE COMPARISON.

Model	Inference Framework	Audio Token	Batch	GPU	Time (s)↓
Qwen2.5-Omni-7B	VLLM [29]	750	Dynamic	1	323
MCAT-Small-9B	Transformer	30	50	4	253
				1	154
MCAT-Large-27B				4	76
				1	337
				4	169

1,000 speech samples using BF16 inference setup with A100 GPUs.

## VI. CONCLUSION

We successfully addressed the critical language scalability and efficiency constraints of MLLMs for the S2TT task. Our primary contributions are twofold: we introduced a novel multilingual S2TT training strategy leveraging curriculum learning for mutual translation across **70** languages (**4,830** directions), and we designed an efficient architecture with an optimized speech adapter that achieved a **25 $\times$**  input compression (reducing tokens from 750 to 30). Crucially, our models (**9B/27B**) surpassed state-of-the-art end-to-end performance on the FLEURS dataset across **70  $\times$  69** directions, despite the extreme compression. This high performance and extensive multilingual support are attained with remarkable resource efficiency, requiring only  $\sim 100M$  trainable parameters and limited data resources (10h S2TT data per language). We confirm that large-scale multilingual S2TT is achievable with minimal computational overhead, proposing a highly scalable and efficient MLLM model.

## LIMITATIONS

This paper presents a method for training an MLLM for languages with less than 10 hours of speech translation data.

However, the performance of S2TT and the range of supported languages are constrained by the capabilities of the LLM. MLLMs trained using this method may not perform well on languages that are not supported by the LLM or on those with poor machine translation performance. Furthermore, for some low-resource languages, additional speech recognition data is still required for initialization.

## REFERENCES

- [1] Y. Du, Y. Pan, Z. Ma, B. Yang, Y. Yang, K. Deng, X. Chen, Y. Xiang, M. Liu, and B. Qin, “Making llms better many-to-many speech-to-text translators with curriculum learning,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 12466–12478. [1](#) [2](#)
- [2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020. [1](#)
- [3] Q. Cheng, M. Fang, Y. Han, J. Huang, and Y. Duan, “Breaking the data barrier: Towards robust speech translation via adversarial stability training,” *arXiv preprint arXiv:1909.11430*, 2019. [1](#)
- [4] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023. [1](#) [2](#)
- [5] M. Sperber and M. Paulik, “Speech translation and the end-to-end promise: Taking stock of where we are,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7409–7421. [1](#)
- [6] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, “Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,” *arXiv preprint arXiv:2305.11000*, 2023. [1](#) [2](#)
- [7] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, “Qwen2-audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024. [1](#) [9](#)
- [8] C. Wang, A. Wu, and J. Pino, “Covost 2 and massively multilingual speech-to-text translation,” *arXiv preprint arXiv:2007.10310*, 2020. [1](#) [2](#) [5](#)
- [9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023. [1](#)
- [10] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742. [1](#) [3](#)
- [11] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech,” *arXiv preprint arXiv:2205.12446*, 2022. [Online]. Available: <https://arxiv.org/abs/2205.12446> [2](#) [5](#)
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518. [2](#) [3](#)
- [13] “Scaling neural machine translation to 200 languages,” *Nature*, vol. 630, no. 8018, pp. 841–846, 2024. [2](#) [6](#) [11](#) [12](#)
- [14] C. Wang, J. Pino, and J. Gu, “Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation,” *arXiv preprint arXiv:2006.05474*, 2020. [2](#)
- [15] “Joint speech and text machine translation for up to 100 languages,” *Nature*, vol. 637, no. 8046, pp. 587–593, 2025. [2](#) [6](#) [7](#) [12](#)
- [16] Y. Li, Z. Liu, Z. Li, X. Zhang, Z. Xu, X. Chen, H. Shi, S. Jiang, X. Wang, J. Wang *et al.*, “Perception, reason, think, and plan: A survey on large multimodal reasoning models,” *arXiv e-prints*, pp. arXiv–2505, 2025. [2](#)
- [17] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, “Salmonn: Towards generic hearing abilities for large language models,” *arXiv preprint arXiv:2310.13289*, 2023. [2](#)
- [18] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy *et al.*, “Voxtoral,” *arXiv preprint arXiv:2507.13264*, 2025. [2](#)
- [19] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, “Qwen3-omni technical report,” *arXiv preprint arXiv:2509.17765*, 2025. [2](#) [5](#) [7](#)
- [20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215. [5](#)
- [21] M. Cui, P. Gao, W. Liu, J. Luan, and B. Wang, “Multilingual machine translation with open large language models at practical scale: An empirical study,” *arXiv preprint arXiv:2502.02481*, 2025. [5](#) [11](#)
- [22] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025. [5](#) [11](#)
- [23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021. [5](#)
- [24] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, “Seamlessm4t-massively multilingual & multimodal machine translation,” *arXiv preprint arXiv:2308.11596*, 2023. [5](#)
- [25] R. Rei, J. G. De Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. Martins, “Comet-22: Unbabel-ist 2022 submission for the metrics shared task,” in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022, pp. 578–585. [5](#) [11](#)
- [26] M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: <https://www.aclweb.org/anthology/W18-6319> [5](#) [11](#)
- [27] Y. Lu, W. Zhu, L. Li, Y. Qiao, and F. Yuan, “Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 10 748–10 772. [6](#) [12](#)
- [28] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, “Qwen2.5-omni technical report,” *arXiv preprint arXiv:2503.20215*, 2025. [7](#)
- [29] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. [9](#)

## APPENDIX

*A. Language Coverage*

The MLLM’s S2TT capability is contingent upon the upper bound of the underlying LLM’s MT performance. Consequently, the MT capability of the base model directly determines the ceiling of our translation quality and guides our final selection of supported languages.

**1) 28 Languages for MCAT-Small:** The GemmaX2-9B [21] model was specifically trained and optimized for these 28 target languages, resulting in a significant performance improvement. Based on this design, these 28 languages are designated as fully supported.

**2) 70 Languages for MCAT-Large:** We used the COMET scoring system and the Flores [13] dataset to evaluate the translation quality of the Gemma3-27B [22] base model. A COMET score of 70 was set as the minimum acceptable threshold. Approximately 70 languages met or exceeded this benchmark, leading to their selection for support.

*B. Evaluation Metrics: COMET vs. spBLEU*

As shown in Table VI and XIII, a notably divergent trend is observed for the average scores in the 9→69 direction between the cascaded NLLB model (21.0/80.5) and MCAT-Large (20.1/81.5). Specifically, NLLB achieves a higher spBLEU score but a lower COMET score compared to MCAT-Large. This phenomenon is rooted in the distinct design philosophies of the models and the metrics: NLLB, as a specialized machine translation model, is optimized for strict sentence-level alignment and high lexical overlap with the reference translations, leading to superior performance on the n-gram-based spBLEU [26]. In contrast, MCAT-Large, an MLLM-based architecture, prioritizes generating fluent and human-natural sentences through flexible paraphrasing and semantic preservation. This semantic quality and fluency, which may come at the expense of rigid word-for-word matching, is better captured by COMET [25], a neural metric that has demonstrated a higher correlation with human judgment of translation quality.

TABLE XII  
SUMMARY OF TRAINING DATASETS FOR MCAT MODELS.

Model	Task	Description	Dataset	Split	Data Size	Metric
	ASR	Automatic Speech Recognition	Common Voice 22 FLEURS	train train	~3500h ~617.7h	WER ↓
MCAT	SMT	Speech-Guided Machine Translation	FLEURS	train	~617.7h	spBLEU / COMET ↑
	SRT	Speech Recognition and Translation	FLEURS	train	~617.7h	spBLEU / COMET ↑

Data size refers to the actual amount used, as we removed overly long samples and balanced the data across different languages.

TABLE XIII  
SPBLEU RESULTS ON 9×27 AND 9×69 DIRECTIONS ON THE FLEURS DATASET.

Systems (X → 27)	ara	cmn	eng	ind	jpn	kor	tha	tur	vie	Avg.
<b>Cascaded ASR+MT Models</b>										
Whisper-Large-V3 + NLLB-200-3.3B [13]	21.5	18.9	30.2	24.2	19.4	19.9	17.5	23.7	18.9	21.6
Whisper-Large-V3 + LLaMAX3-8B-Alpaca [27]	17.5	16.1	25.2	20.6	16.1	16.9	14.5	19.0	16.2	18.0
<b>End-to-end S2TT Models</b>										
SeamlessM4T-V2-Large [15]	15.5	13.2	30.9	15.3	12.2	14.4	11.5	15.7	13.9	15.8
MCAT-Small-9B (ours)	20.2	18.4	32.7	24.5	18.9	21.4	16.8	23.5	18.8	21.7
MCAT-Large-27B (ours)	20.2	17.9	31.7	24.4	18.3	21.1	15.7	24.2	18.2	21.3
Systems (X → 69)	ara	cmn	eng	ind	jpn	kor	tha	tur	vie	Avg.
<b>Cascaded ASR+MT Models</b>										
Whisper-Large-V3 + NLLB-200-3.3B [13]	21.0	18.0	30.1	23.4	18.5	19.3	16.8	23.1	18.4	21.0
Whisper-Large-V3 + LLaMAX3-8B-Alpaca [27]	15.7	14.4	23.5	18.9	14.2	15.1	12.6	17.2	14.6	16.2
<b>End-to-end S2TT Models</b>										
SeamlessM4T-V2-Large [15]	15.8	13.5	31.7	15.4	11.9	14.3	11.7	16.0	14.3	16.1
MCAT-Large-27B (ours)	19.0	16.8	30.3	23.0	17.1	19.7	14.8	22.8	17.1	20.1