MITIGATING THE CURSE OF MULTILINGUALITY IN MANY-TO-MANY SPEECH-TO-TEXT TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have achieved significant success in speech-to-text translation (S2TT). However, when processing multilingual speech inputs, a single speech encoder shared across all languages suffers from the curse of multilinguality: languages at different resource levels compete for limited representation capacity, leading to strong high-resource performance but substantial degradation on low-resource speech. To address this problem and improve multilingual consistency, we propose MSRT, a novel framework built around a resource-aware Mixture of Speech Encoders (MoSE). MoSE uses an explicit language router to assign each utterance to an appropriate expert encoder. A frozen expert preserves high-resource language capabilities, while a trainable expert adapts to and specializes in medium- and low-resource languages. We further introduce a five-stage curriculum learning strategy that substantially reduces data dependence, requiring under 10 hours of paired S2TT data per language on average for effective alignment. We conduct extensive experiments on 45 languages, systematically evaluating all 45×44 translation directions. Our models, spanning 4B and 12B parameters, achieve state-of-the-art performance, outperforming substantially larger open-source models. Empirical analyses show that MoSE improves high-, medium-, and low-resource languages simultaneously, with the largest gains on low-resource speech, thereby mitigating the curse of multilinguality without compromising high-resource performance. To support future multilingual research, we will release our code and models.

1 INTRODUCTION

Speech-to-text translation (S2TT) aims to translate speech from a source language into text in a target language. Traditionally, these systems have utilized cascaded pipelines: an Automatic Speech Recognition (ASR) component first extracts a transcript from the speech (Radford et al., 2023), which a subsequent Machine Translation (MT) model then converts into the desired language (Shang et al., 2026). Recently, there has been a growing focus on integrating speech encoders with large language models (LLMs) to endow them with multilingual speech capabilities, thereby advancing ASR (Ma et al., 2026) and S2TT tasks (Du et al., 2026).

Despite these advances, current MLLMs remain strongly English-centric in S2TT. They perform well on English-to-X directions but degrade in many-to-many translation from non-English speech. As shown in Figure 1, their consistency scores for low-resource speech inputs fall substantially below those for English, revealing a pronounced cross-lingual performance gap. This disparity is not merely a target-generation issue: changing the source speech language causes substantial quality variation even when the target set and decoder remain unchanged. Consequently, English-centric results can overstate multilingual coverage and obscure failures on multilingual speech inputs.

One important cause of this inconsistency is that most multilingual MLLMs (Xu et al., 2025) rely on a single shared speech encoder for all input languages. As language coverage expands, this architecture suffers from the "**curse of multilinguality**" (Conneau et al., 2020): limited encoder capacity forces languages with different levels of available resources to compete for the same representation space. High-resource languages require preserving pretrained acoustic knowledge, whereas medium- and low-resource languages require greater adaptation. Joint optimization can therefore overwrite established representations or underfit scarce-language features.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

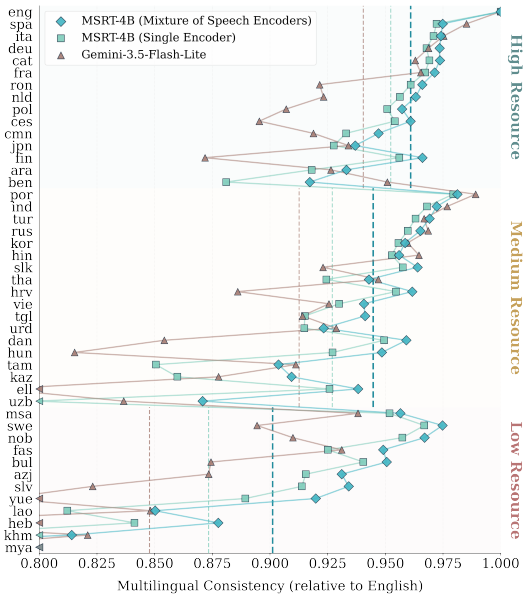


Figure 1: Multilingual consistency by source language and resource level. Consistency is the ratio of a source language's average translation score (across 44 target directions) to the English average. Higher is better, and dashed vertical lines indicate each model's average within a resource group.

Inspired by the Mixture-of-Experts architectures in LLMs (Blevins et al., 2024), we propose MSRT, an S2TT MLLM built on a resource-aware **Mixture of Speech Encoders (MoSE)**: two homogeneous Whisper encoders (Radford et al., 2023) with asymmetric roles, where a frozen expert handles high-resource languages and a trainable expert adapts to medium- and low-resource speech. Source- and target-language tokens route each utterance to one expert and control generation, specializing at the inference cost of a single encoder. We further extend the curriculum strategy (Du et al., 2025) into five stages of increasing difficulty.

We evaluate MSRT on FLEURS across all 45×44 non-identity directions, comparing representative languages from 11 families in detail. MSRT-4B and MSRT-12B achieve the highest average COMET scores of 82.5 and 83.8, outperforming the Gemini-3.5-Flash-Lite API and the substantially larger MCAT-27B (Du et al., 2026). Across the full direction grid, they reach at least 80 COMET on 1,552 and 1,706 of 1,980 directions, against 928 for Gemini. Ablation studies show that MoSE improves high-, medium-, and low-resource languages alike, with the largest gains on low-resource speech, and it transfers to CoVoST-2. These gains indicate that MoSE mitigates the curse of multilinguality rather than trading high-resource performance for low-resource improvements. Our contributions are summarized as follows:

- We propose **MoSE**, a resource-aware mixture-of-speech-encoder architecture that **mitigates the curse of multilinguality** by preserving high-resource knowledge while enabling **targeted adaptation for medium- and low-resource languages**.
- We introduce **MSRT**, a model family of **4B and 12B** parameters, both of which support **all translation directions among 45 languages** using **under 10 hours** of paired S2TT data per language on average.
- We comprehensively evaluate **all 1,980 translation directions across 45 languages**, demonstrating broad and reliable many-to-many translation across diverse language families and resource levels, especially for **low-resource speech** in realistic multilingual settings. We will release our code and models to facilitate reproducible research.

2 Related Work

Speech-to-Text Translation. Conventional speech-to-text translation (S2TT) cascades automatic speech recognition and machine translation. Although it can reuse strong task-specific models, recognition errors propagate downstream and two-stage decoding increases latency. End-to-end S2TT instead learns a direct mapping from source speech to target text. SeamlessM4T integrates multilingual speech recognition and translation in a single model but still depends heavily on paired supervision (Barrault et al., 2023). ZeroSwot reduces this dependence by connecting an ASR-trained speech encoder to a massively multilingual translation model through CTC compression and optimal-transport alignment (Tsiamas et al., 2024), making speech-text alignment a central challenge for end-to-end translation.

Many-to-Many Speech-to-Text Translation. Many-to-many S2TT expands both source- and target-language coverage, requiring one model to preserve language-specific acoustic cues while controlling multilingual generation. General-purpose multimodal LLMs such as Qwen3-Omni-30B-Instruct show that pretrained language backbones can support multilingual speech understanding and translation (Xu et al., 2025). Specialized MLLMs more directly exploit the translation knowledge of LLMs: LLM-SRT transfers it to many-to-many S2TT through curriculum learning (Du et al., 2025), while MCAT scales MLLM-based translation to 70 languages (Du et al., 2026). Nevertheless, increasing language coverage does not eliminate resource imbalance or interference between languages. This limitation motivates speech-side mechanisms that allocate capacity according to each language's data resources and acoustic variation.

Mixture of Experts. MoE has been explored in multimodal and speech models from several perspectives. Uni-MoE activates modality-specific experts to reduce cross-modal interference in a unified MLLM (Li et al., 2025), whereas SimulMega uses expert routers as policies for balancing latency and quality in simultaneous translation (Le et al., 2025). PaM selects heterogeneous audio encoders from task prompts to serve different audio tasks (Shan et al., 2025), while DialectMoE learns dynamic expert routing for multi-dialect ASR (Zhou et al., 2024). Closest to us, LCMA-SRT routes with language-conditional MoE adapters over a shared transducer (Li et al., 2026); MoSE instead splits whole speech encoders by resource level and freezes one to protect high-resource languages.

3 Methodology

3.1 Model Architecture

As shown in Figure 2, MSRT connects a resource-aware Mixture of Speech Encoders (MoSE), a speech adapter, and a pretrained LLM.

Mixture of Speech Encoders. MoSE comprises two encoders and an explicit language router. Instead of sharing one encoder across languages with unequal supervision, it separates high-resource knowledge preservation from medium- and low-resource adaptation while activating one expert per utterance. This design mitigates the curse of multilinguality by improving underrepresented languages without sacrificing high-resource performance.

- **Frozen Expert.** The frozen expert E_f retains the original Whisper encoder with fixed parameters. It preserves general speech representations for high-resource languages, prevents catastrophic forgetting, and anchors the shared speech interface during adaptation.
- **Trainable Expert.** The trainable expert E_t shares the Whisper initialization but adapts to medium- and low-resource speech, capturing underrepresented accents and phonological patterns. Its architecture matches E_f, so both experts produce compatible representations and share one adapter without expert-specific projections.
- **Explicit Language Router.** The router maps each source-language token to its resource-group expert: high-resource languages use E_f, whereas medium- and low-resource languages use E_t. Table 2 lists the resource group of each of the 45 languages. Utterance-level routing keeps all frames on one acoustic path and requires neither a learned gate nor a balancing loss.

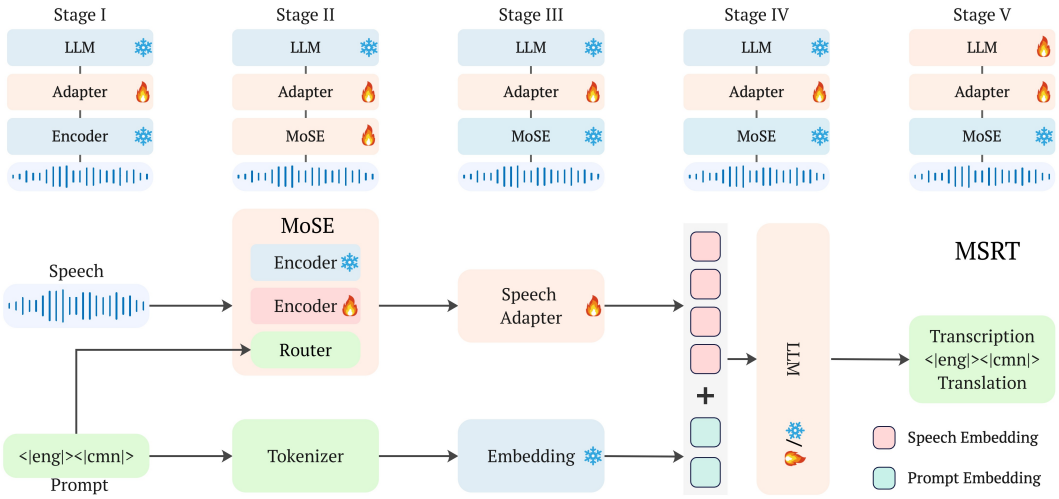


Figure 2: Overview of the MSRT framework. Based on the source-language token in the prompt, MoSE routes high-resource speech to a frozen encoder and medium- and low-resource speech to a trainable encoder.

Overall, let $\mathcal{R}(\ell_s)$ denote the expert selected for source language ℓ_s. The MoSE output is written as

$$H_{\text{MoSE}}(x, \ell_s) = E_{\mathcal{R}(\ell_s)}(x), \quad \mathcal{R}(\ell_s) \in \{f, t\}. \tag{1}$$

Here, in Equation 1, $\mathcal{R}(\ell_s) = f$ for $\ell_s \in \mathcal{L}_{\text{high}}$ and $\mathcal{R}(\ell_s) = t$ otherwise, covering the medium- and low-resource groups. This formulation retains the inference cost of a single encoder.

Speech Adapter. The selected encoder produces variable-length features. A Q-Former compresses them into a translation-aware representation, while an MLP projects it into the LLM hidden space, removing acoustic redundancy while retaining information required for translation.

LLM. The pretrained LLM provides multilingual translation knowledge and remains frozen during speech-side alignment to prevent catastrophic forgetting. It is unlocked only in the final training stage, where it adapts to the aligned acoustic representation.

Tokenizer Expansion. We add dedicated language tokens for the 45 FLEURS languages we cover (Conneau et al., 2023). The source-language token controls MoSE encoder selection, while the target-language token specifies the LLM translation direction. Together, they provide explicit language control and reduce unintended switching.

Text Prompt. As shown in Figure 2 and Table 1, the prompt explicitly instructs the LLM to translate from the source language into the target language.

Table 1: Prompt design for the five-stage curriculum.

Stage	Task	Speech	Text Prompt	Prediction
I	ASR	✓	<|eng|>	Transcription
II	ASR	✓	<|eng|>	Transcription
III	SMT	✓	Transcription<|eng|><|cmn|>	Translation
IV	SRT	✓	<|eng|><|cmn|>	Transcription<|eng|><|cmn|> Translation
V	SRT	✓	<|eng|><|cmn|>	Transcription<|eng|><|cmn|> Translation

3.2 Five-Stage Curriculum Learning

Following Du et al. (2025), we organize ASR, speech-guided machine translation (SMT), and joint speech recognition and translation (SRT) into the five stages in Table 1; Table 5 summarizes the modules optimized per stage, and each stage starts from the preceding checkpoint.

- **Stage I: ASR alignment.** The Q-Former and MLP adapter are trained successively on English-only, high-resource multilingual, and fully multilingual ASR data, establishing a stable speech-text interface while both experts and the LLM remain frozen.
- **Stage II: Expert specialization.** MoSE routing is activated for multilingual ASR. The frozen expert E_f preserves high-resource representations, whereas E_t is optimized for medium- and low-resource speech; the shared adapter continues training.
- **Stage III: Translation activation.** SMT supplies the ground-truth transcription together with source- and target-language tokens. The adapter is optimized to connect routed speech features to the translation knowledge of the frozen LLM.
- **Stage IV: End-to-end SRT.** The ground-truth transcription is removed from the input. With the experts and LLM fixed, the adapter learns to generate both the transcription and translation directly from routed speech.
- **Stage V: Joint adaptation.** LLM LoRA is enabled and jointly optimized with the adapter for SRT, adapting multilingual generation to acoustic inputs without updating the pretrained LLM weights, so that the whole pipeline is optimized end to end.

4 Experimental Settings

Datasets. We use Common Voice 24 (Ardila et al., 2020) for multilingual ASR pre-training. For S2TT training, we use paired speech-text data from FLEURS, comprising 5.9–12.1 hours of speech per language and under 10 hours on average. We evaluate all systems on FLEURS. We additionally use the CoVoST-2 (Wang et al., 2021) benchmark for the data-scaling study.

Training Details. MSRT pairs a pretrained multilingual LLM (Shang et al., 2026) with MoSE and a speech adapter; Table 5 in the appendix lists every component and the stage at which it is optimized. Training uses BF16, DeepSpeed ZeRO-0, and AdamW with a peak learning rate of 1×10^{-4} and 1,000 warmup steps, for three days on eight Ascend 910C NPUs.

Baselines. We compare three representative deployment paradigms:

- **API-based:** Gemini-3.5-Flash-Lite (Google DeepMind, 2026) serves as the API baseline.
- **Cascade:** Whisper-Large-v3 (Radford et al., 2023) first transcribes speech, after which NLLB-200-3.3B (NLLB Team et al., 2024) translates the transcription.
- **End-to-end:** SeamlessM4T-V2-Large (Barrault et al., 2023), Qwen3-Omni-30B-Instruct (Xu et al., 2025), and MCAT-27B (Du et al., 2026) translate speech directly.

Evaluation Metrics. We employ COMET (Rei et al., 2022) and spBLEU (Post, 2018), the latter with the FLORES-200 tokenizer; all results use beam size 1.

Table 2: Language support. Resource levels follow SeamlessM4T (Barrault et al., 2023).

Resource	Languages (ISO Code)
High	**Arabic (ara)**, Bengali (ben), Catalan (cat), Czech (ces), **Chinese (cmn)**, German (deu), **English (eng)**, Finnish (fin), French (fra), Italian (ita), **Japanese (jpn)**, Dutch (nld), Polish (pol), Romanian (ron), Spanish (spa)
Medium	Danish (dan), Greek (ell), Hindi (hin), Croatian (hrv), **Hungarian (hun)**, **Indonesian (ind)**, Kazakh (kaz), **Korean (kor)**, Portuguese (por), Russian (rus), Slovak (slk), **Tamil (tam)**, Tagalog (tgl), **Thai (tha)**, **Turkish (tur)**, Urdu (urd), Uzbek (uzb), **Vietnamese (vie)**
Low	Azerbaijani (azj), Bulgarian (bul), Persian (fas), Hebrew (heb), Khmer (khm), Lao (lao), Malay (msa), Burmese (mya), Norwegian (nob), Slovenian (slv), Swedish (swe), Cantonese (yue)
Total	**45 languages. Bold** marks the 11 main-evaluation languages, one from each of 11 families.

Table 3: spBLEU / COMET on FLEURS for 11×44, 44×11, and 45×44 averages. The 11 languages cover 11 distinct language families, with one language selected from each family. The full $45 \rightarrow 44$ grid is reported per source language in Table 6 and per target language in Table 7.

Direction	Gemini-3.5 -Flash-Lite	Whisper+ NLLB-3.3B	SeamlessM4T -V2-Large	Qwen3-Omni -30B-Instruct	MCAT -27B	MSRT-4B (ours)	MSRT-12B (ours)
ara→44	22.2 / 80.2	21.1 / 78.7	15.7 / 70.1	20.5 / 78.8	20.0 / 79.7	23.7 / 82.0	**26.5 / 84.1**
cmn→44	17.2 / 79.5	18.5 / 80.5	13.4 / 74.0	20.9 / 82.8	17.7 / 81.3	20.8 / 83.2	**22.2 / 84.2**
eng→44	34.2 / 86.6	30.5 / 84.3	31.8 / 85.3	31.9 / 85.7	31.9 / 87.1	34.2 / 87.9	**36.1 / 88.7**
hun→44	14.8 / 70.6	21.2 / 79.9	15.3 / 68.9	11.0 / 65.0	19.6 / 79.9	23.8 / 83.4	**26.1 / 85.0**
ind→44	27.4 / 84.6	23.7 / 82.2	15.2 / 71.5	24.7 / 83.4	24.2 / 84.2	27.3 / 85.4	**28.7 / 86.4**
jpn→44	18.4 / 80.8	18.9 / 80.5	11.9 / 69.5	19.5 / 81.8	18.0 / 80.8	20.4 / 82.3	**22.0 / 84.0**
kor→44	20.8 / 83.1	19.5 / 81.8	14.2 / 74.1	20.4 / 82.7	20.9 / 83.4	22.4 / 84.2	**24.4 / 85.6**
tam→44	17.3 / 78.9	12.9 / 71.4	12.6 / 69.8	2.5 / 50.4	13.0 / 74.0	17.9 / 79.4	**20.3 / 81.8**
tha→44	20.2 / 82.0	17.1 / 78.6	11.5 / 69.0	18.8 / 81.3	15.6 / 78.9	20.8 / 82.9	**22.4 / 83.9**
tur→44	25.6 / 83.7	23.4 / 83.0	15.8 / 72.1	23.2 / 82.5	24.0 / 84.1	25.7 / 85.2	**28.2 / 86.4**
vie→44	19.7 / 80.1	18.7 / 78.9	14.1 / 72.6	20.0 / 81.3	18.0 / 80.5	21.4 / 82.7	**23.3 / 84.1**
Avg. 11→44	21.6 / 80.9	20.5 / 80.0	15.6 / 72.4	19.4 / 77.8	20.3 / 81.3	23.5 / 83.5	**25.5 / 84.9**
44→ara	18.6 / 74.1	20.9 / 75.8	16.8 / 71.2	19.1 / 73.7	21.1 / 78.4	23.2 / 80.5	**25.3 / 81.7**
44→cmn	21.1 / 78.3	14.3 / 71.2	11.3 / 65.4	21.3 / 78.4	21.5 / 80.2	25.2 / 83.3	**27.4 / 84.5**
44→eng	24.9 / 79.2	31.7 / 80.5	31.1 / 83.3	28.1 / 79.7	27.5 / 80.6	34.0 / 84.1	**35.6 / 85.2**
44→hun	20.6 / 79.7	18.4 / 75.7	13.8 / 67.9	15.9 / 74.4	17.8 / 79.7	21.4 / 82.4	**23.3 / 84.2**
44→ind	21.7 / 80.9	22.9 / 80.5	16.1 / 76.7	21.0 / 79.0	21.9 / 82.1	25.3 / 85.0	**27.0 / 86.1**
44→jpn	18.2 / 81.8	9.3 / 76.1	7.3 / 71.9	17.2 / 81.0	18.2 / 83.1	19.9 / 85.4	**22.1 / 86.6**
44→kor	16.0 / 79.9	13.9 / 77.5	10.2 / 71.8	15.2 / 78.7	15.5 / 80.7	16.8 / 83.2	**18.8 / 84.3**
44→tam	20.0 / 80.0	19.8 / 78.5	17.6 / 75.9	15.1 / 74.7	19.0 / 81.0	21.2 / 83.1	**23.7 / 84.2**
44→tha	27.6 / 78.1	20.2 / 74.0	20.1 / 70.3	26.7 / 78.1	26.2 / 79.8	29.9 / 82.5	**31.8 / 83.6**
44→tur	19.6 / 78.4	18.3 / 76.0	13.6 / 70.0	15.8 / 73.8	18.6 / 78.9	21.1 / 81.6	**23.1 / 83.4**
44→vie	23.5 / 79.4	24.0 / 78.7	18.4 / 73.8	22.2 / 77.3	23.2 / 81.1	27.1 / 83.9	**28.4 / 84.8**
Avg. 44→11	21.1 / 79.1	19.4 / 76.8	16.0 / 72.6	19.8 / 77.2	21.0 / 80.5	24.1 / 83.2	**26.1 / 84.4**
Avg. 45→44	20.3 / 78.3	19.4 / 76.5	15.3 / 71.5	17.3 / 74.1	19.7 / 79.6	23.2 / 82.5	**25.0 / 83.8**

Underlining denotes the second-best results, while **bold blue** values indicate the best performance.

5 Experiments

5.1 Main Results

Overall Analysis. On the overall $45 \rightarrow 44$ average of Table 3, MSRT-12B and MSRT-4B lead with 83.8 and 82.5 COMET, ahead of MCAT-27B (79.6), Gemini-3.5-Flash-Lite (78.3), Whisper+NLLB-3.3B (76.5), Qwen3-Omni-30B-Instruct (74.1), and SeamlessM4T-V2-Large (71.5). spBLEU ranks them identically: 25.0 and 23.2 against 20.3 and 19.7 for Gemini-3.5-Flash-Lite and the larger MCAT-27B. They also lead the 11×44 and 44×11 subsets and English→44 (88.7 and 87.9; Table 8), by 4.2 COMET on the full grid and 1.6 on English. MSRT-12B and MSRT-4B are the two best systems on COMET in every row of Table 3.

English-centric vs. Many-to-Many Translation. Although optimized for many-to-many translation, MSRT is also strong on English-source directions, averaging 87.9 and 88.7 COMET on eng→44 against 82.5 and 83.8 on the full $45 \rightarrow 44$ grid. Even MSRT-4B, one seventh of MCAT-27B's size, edges past it here (87.9 vs. 87.1). The gap between the two settings is informative: SeamlessM4T-V2-Large loses 13.8 COMET once non-English sources are included and Qwen3-Omni-30B-Instruct loses 11.6, whereas MSRT-4B and MSRT-12B lose only 5.4 and 4.9. MSRT-12B retains 94.5% of that English-source COMET, against 83.8% for SeamlessM4T-V2-Large.

High-Resource vs. Low-Resource Languages. Table 3 shows Qwen3-Omni-30B-Instruct degrading sharply on lower-resource sources: 2.5/50.4 spBLEU/COMET on Tamil (tam) and 11.0/65.0 on Hungarian (hun), far below its 31.9/85.7 on English, though 44→tam still reaches 15.1/74.7. MSRT-4B and MSRT-12B stay strong in both roles, because MoSE's trainable expert specializes in medium- and low-resource languages while the frozen expert preserves high-resource quality. Over the 12 low-resource sources, MSRT-12B averages 80.3 COMET against 75.0 for MCAT-27B. MSRT-12B reaches 84.2 COMET on both 44→tam and 44→hun, against 74.7 and 74.4 for Qwen3-Omni-30B-Instruct.

Table 4: Distribution of the 1,980 translation directions across COMET bands.

COMET	Gemini-3.5 -Flash-Lite	Whisper+ NLLB-3.3B	SeamlessM4T -V2-Large	Qwen3-Omni -30B-Instruct	MCAT -27B	MSRT-4B (ours)	MSRT-12B (ours)
≥ 80	928	1038	117	881	1232	<u>1552</u>	**1706**
$[70, 80)$	815	680	1109	557	554	<u>359</u>	**216**
< 70	237	262	754	542	194	<u>69</u>	**58**

<u>Underlining</u> and **bold blue** mark the second-best and best counts; the two lower bands are smaller-is-better.

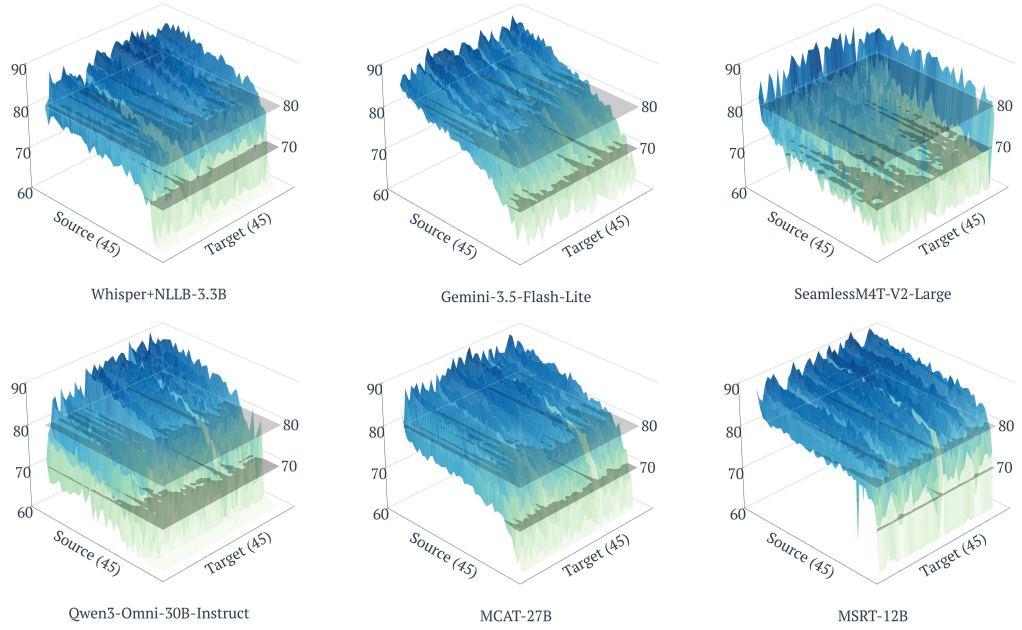


Figure 3: COMET over the full source–target grid. Shaded regions highlight scores below 80 and 70; the diagonal is smoothed for visualization and excluded from every statistic reported in the text.

Analysis of the Landscape Figure. Figure 3 gives the full picture over the grid. Every model peaks on English-source speech, so the peak height is uninformative; what separates them is the height of the remaining surface. MSRT-12B averages at least 80 COMET for 41 of the 45 source languages and MCAT-27B for 29, while SeamlessM4T-V2-Large is an isolated ridge above 80 for English alone.

Source vs. Target Symmetry. Figure 3 also shows whether a language is treated consistently in its two roles. Most baselines are not: Whisper+NLLB-3.3B averages 49.4 COMET more on Burmese as a target than as a source, Qwen3-Omni-30B-Instruct loses 25.7 on Kazakh and 24.3 on Tamil as source, and Gemini-3.5-Flash-Lite differs by more than 5 COMET on 23 of the 45 languages.

Scaling the Backbone. MoSE is architecture-agnostic: we train MSRT-12B by replacing the 4B backbone with a 12B LLM while keeping the routing policy, the two expert encoders, and the five-stage curriculum unchanged. The overall $45\rightarrow44$ COMET average rises from 82.5 to 83.8 and the Eng$\rightarrow$44 average from 87.9 to 88.7 (Table 3). The gain is broad: all 44 English-source directions improve on COMET, and source languages averaging above 85 rise from 10 to 23.

Parameter Efficiency. With only 4B/12B parameters, MSRT surpasses the larger Qwen3-Omni-30B-Instruct and MCAT-27B, reaching COMET ≥ 80 on 1,552 and 1,706 of the 1,980 directions against 881 and 1,232 (Table 4); directions below 70 COMET shrink from 69 to 58. The advantage holds source language by source language in Figure 3: MSRT-12B beats MCAT-27B, Qwen3-Omni-30B-Instruct, and Whisper+NLLB-3.3B on all 45 source languages, and Gemini-3.5-Flash-Lite and SeamlessM4T-V2-Large on 44 of them.

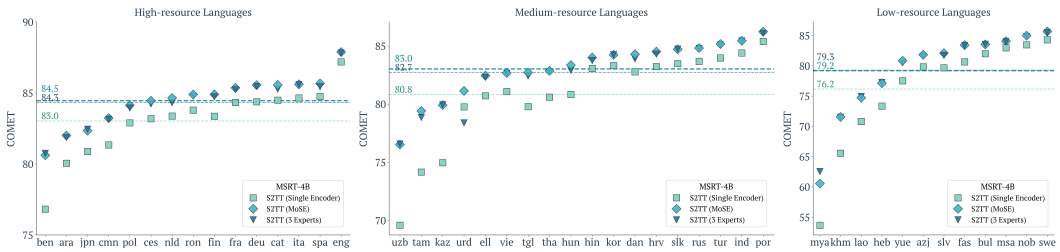


Figure 4: Expert-count ablation across resource levels: S2TT with one shared encoder, with the resource-aware two-expert split (MoSE), and with three experts. Consistent gains, largest for low-resource languages, show that MoSE mitigates the curse of multilinguality, while a third expert moves the overall average by only 0.2 COMET.

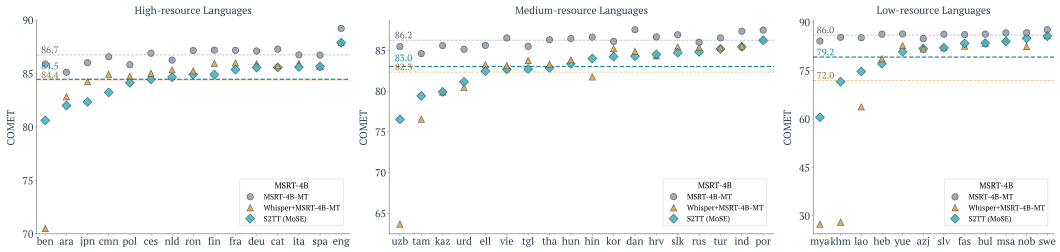


Figure 5: S2TT, MSRT-4B-MT, and the Whisper+MSRT-4B-MT cascade across resource levels. MSRT-4B-MT translates gold transcriptions and therefore upper-bounds S2TT. Cascading Whisper-Large-v3 into MSRT-4B-MT matches end-to-end S2TT on high-resource speech but loses ground as source resources shrink.

5.2 Ablation Study and Analysis

Ablation of MoSE. Figure 4 contrasts a single shared encoder with the resource-aware two-expert split. Removing the split costs 2.2 COMET overall (82.5 to 80.3), and the damage is universal: the split wins on all 45 source languages, from 0.7 on English to 6.9 on Uzbek and Burmese. Its five largest gaps—Uzbek and Burmese (6.9 each), Khmer (6.0), Tamil (5.3), Kazakh (4.9)—include three medium-resource languages, so those with the least data lose their dedicated representation first.

Ablation of Expert Count. A third expert, one per resource group, does not pay off (Figure 4). The overall average moves from 82.5 to 82.3 COMET, the three-expert variant is better by more than 0.1 COMET on only 5 source languages and worse by more than 0.1 on 23, and every group average stays within 0.3 COMET of the two-expert split. Its largest win (2.0 on Burmese) is offset by its largest loss (2.7 on Urdu). Two experts therefore suffice at this scale: the third adds an encoder and changes nothing measurable.

Ablation of Resource Level. The ordering of the gains matters: 3.0 COMET for low-resource sources, 2.2 for medium-resource, and 1.4 for high-resource, all positive (Figure 4). Multilingual systems usually buy low-resource quality by giving up high-resource quality; here every group improves and the largest gain falls on the group with the least data. With the expert-count result, this rules out raw capacity: what matters is where the capacity is placed, not how much of it there is.

Comparison with Text MT and Cascaded Systems. We isolate the MT stage with **MSRT-4B-MT**, the SMT variant of the same curriculum run on ground-truth transcriptions, and its Whisper-Large-v3 cascade **Whisper+MSRT-4B-MT**. MSRT-4B-MT upper-bounds S2TT; its distance to MSRT-4B locates the remaining headroom (Figure 5): 2.3 COMET on high-resource languages but 3.2 and 6.8 on medium- and low-resource ones. The cascade points the same way: Whisper+MSRT-4B-MT matches S2TT on high-resource speech but trails by 0.7 and 7.2 points elsewhere. S2TT is therefore limited by speech representation and cross-modal alignment rather than by the MT stage.

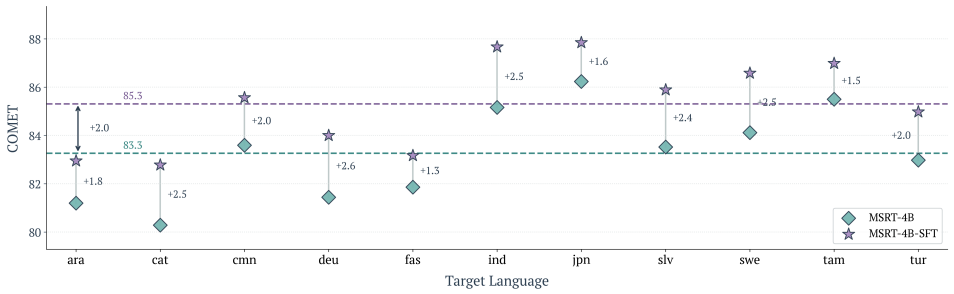


Figure 6: Data scaling on CoVoST-2 English-to-X directions.

Data Scaling. Figure 6 evaluates zero-shot transfer to 11 CoVoST-2 English-to-X directions: MSRT-4B, whose English-side data is the **7.5 h** of FLEURS English speech it is trained on, reaches an average of 83.3 COMET with every direction clearing 80. We additionally continue MSRT-4B on **429.6 h** of CoVoST-2 English speech to obtain **MSRT-4B-SFT**, which lifts the average to **85.3**, with every direction gaining between 1.3 and 2.6 points. The two systems differ in both the English-side budget and the source-domain match: MSRT-4B is trained only on FLEURS English and never sees CoVoST-2 speech, whereas MSRT-4B-SFT is continued on it. The 83.3 average is therefore a cross-dataset result, while the 85.3 average combines in-domain adaptation with additional data and thus upper-bounds what scaling alone would yield rather than isolating it.

6 Conclusion

We introduced MSRT, a many-to-many S2TT framework instantiated at 4B and 12B parameters that combines resource-aware MoSE with five-stage curriculum learning. By separating high-resource preservation from medium- and low-resource adaptation, MoSE improves all resource groups, with the largest gains on low-resource speech, and the gains persist as the backbone is scaled up. Evaluation of all 1,980 directions among 45 languages demonstrates strong multilingual consistency, parameter efficiency, and cross-dataset generalization, showing that MoSE effectively mitigates the curse of multilinguality. Future work will extend MoSE to broader language coverage.

Limitations

MSRT derives its S2TT capability from the pretrained LLM, so its quality is bounded by that LLM's MT performance on weakly represented languages and directions. It also requires the source language as an explicit input and performs no language identification, so mislabeled or code-switched audio can be routed to the wrong expert.

Reproducibility Statement

Section 3 describes MoSE, the expert-routing policy, and the five-stage curriculum, and Section 4 specifies the datasets, the 45-language inventory with each language's resource group (Table 2), the training setup, the baselines, and the evaluation metrics. The appendix lists the module configuration of every component (Table 5) and reports spBLEU and COMET for all source and target languages (Tables 6–8), so that the aggregate scores in the main tables and figures can be traced back to individual translation directions. All models are scored with the same metrics, and the resource groups that drive expert routing follow SeamlessM4T (Barrault et al., 2023). We will release the code and models used in this work.

AI Use Statement

LLMs were used only for grammar and readability checking of the manuscript.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/. 5

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. SeamlessM4T: Massively multilingual and multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023. 3, 5, 9

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10822–10837, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.604. URL https://aclanthology.org/2024.emnlp-main.604/. 2

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020. 1

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop*, pp. 798–805. IEEE, 2023. 4

Yexing Du, Youcheng Pan, Ziyang Ma, Bo Yang, Yifan Yang, Keqi Deng, Xie Chen, Yang Xiang, Ming Liu, and Bing 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)*, pp. 12466–12478, 2025. 2, 3, 5

Yexing Du, Kaiyuan Liu, Youcheng Pan, Bo Yang, Keqi Deng, Xie Chen, Yang Xiang, Ming Liu, Bing Qin, and Yaowei Wang. MCAT: Scaling many-to-many speech-to-text translation with MLLMs to 70 languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2026. 1, 2, 3, 5

Google DeepMind. Gemini 3.5 Flash Lite. Gemini API model documentation, 2026. URL https://ai.google.dev/gemini-api/docs/models. Accessed: 2026-09-25. 5

Chenyang Le, Bing Han, Jinshun Li, Songyong Chen, and Yanmin Qian. Simulmega: Moe routers are advanced policy makers for simultaneous speech translation. *Advances in Neural Information Processing Systems*, 38:42739–42761, 2025. 3

Nanjie Li, Xiaoyong Guo, Hao Huang, Haihua Xu, and Wei Shi. LCMA-SRT: Language-conditional mixture-of-experts adapters for joint multilingual speech recognition and translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35363–35377, San Diego, California, United States, July 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.acl-long.1634/. 3

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439, 2025. doi: 10.1109/TPAMI.2025.3532688. 3

Ziyang Ma, Guanrou Yang, Wenxi Chen, Zhifu Gao, Yexing Du, Xiquan Li, Zhisheng Zheng, Haina Zhu, Jianheng Zhuo, Zheshu Song, Ruiyang Xu, Tiranrui Wang, Yifan Yang, Yanqiao Zhu, Zhikang Niu, Liumeng Xue, Yinghao Ma, Ruibin Yuan, Shiliang Zhang, Kai Yu, Eng Siong Chng, and Xie Chen. Slam-llm: A modular, open-source multimodal large language model framework and best practice for speech, language, audio and music processing. *IEEE Journal of Selected Topics in Signal Processing*, 20(1):63–76, 2026. doi: 10.1109/JSTSP.2026.3653157. 1

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. 5

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*, pp. 186–191, 2018. 5

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518, 2023. 1, 2, 5

Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pp. 578–585, 2022. 5

Weiqiao Shan, Yuang Li, Yuhao Zhang, Yingfeng Luo, Chen Xu, Xiaofeng Zhao, Long Meng, Yunfei Lu, Min Zhang, Hao Yang, Tong Xiao, and JingBo Zhu. Enhancing speech large language models with prompt-aware mixture of audio encoders. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19305–19320, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.974. URL https://aclanthology.org/2025.emnlp-main.974/. 3

Yuzhe Shang, Pengzhi Gao, Wei Liu, Jian Luan, and Jinsong Su. Scaling model and data for multilingual machine translation with open large language models, 2026. 1, 5

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. Pushing the limits of zero-shot end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14245–14267, 2024. 3

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. CoVoST 2 and massively multilingual speech-to-text translation. In *Proceedings of Interspeech*, pp. 2247–2251, 2021. 5

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1, 3, 5

Jie Zhou, Shengxiang Gao, Zhengtao Yu, Ling Dong, and Wenjun Wang. DialectMoE: An end-to-end multi-dialect speech recognition model with mixture-of-experts. In Sun Maosong, Liang Jiye, Han Xianpei, Liu Zhiyuan, and He Yulan (eds.), *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1148–1159, Taiyuan, China, July 2024. Chinese Information Processing Society of China. URL https://aclanthology.org/2024.ccl-1.89/. 3

A Experimental Details and Full Results

Table 5: MLLM settings for MSRT. The Q-Former uses 80 queries of hidden size 768.

Component	Modules	Stage	Details
Speech Encoder (MoSE)	Frozen Expert (E_f)	–	For high-resource
	Trainable Expert (E_t)	II	For medium-/low-resource
	Language Router	–	Explicit expert routing
Speech Adapter	Q-Former	I–V	Feature compression
	MLP	I–V	Dimension alignment
LLM	+ Vocabulary	–	Language tokens
	+ LoRA	V	$r = 16$, $\alpha = 32$

Table 6: spBLEU / COMET results on $X \rightarrow 44$ directions.

ISO	Gemini-3.5 -Flash-Lite	Whisper+ NLLB-3.3B	SeamlessM4T -V2-Large	Qwen3-Omni -30B-Instruct	MCAT -27B	MSRT-4B (ours)	MSRT-12B (ours)
ara	22.2 / 80.2	21.1 / 78.7	15.7 / 70.1	20.5 / 78.8	20.0 / 79.7	23.7 / 82.0	**26.5 / 84.1**
azj	14.7 / 75.6	16.9 / 78.2	13.6 / 73.5	7.7 / 65.2	15.6 / 78.8	18.7 / 81.8	**21.2 / 83.9**
ben	20.6 / 82.3	8.0 / 60.5	14.6 / 73.9	11.6 / 70.9	14.8 / 76.6	18.7 / 80.6	**21.1 / 82.7**
bul	20.1 / 75.7	22.8 / 80.1	18.6 / 73.9	18.4 / 75.6	23.1 / 82.0	26.0 / 83.5	**28.2 / 85.1**
cat	27.8 / 83.3	26.2 / 82.9	22.2 / 77.9	21.7 / 78.3	24.9 / 83.7	28.5 / 85.6	**29.7 / 86.4**
ces	20.7 / 77.5	22.5 / 81.0	18.1 / 73.7	19.1 / 77.2	22.7 / 82.8	25.9 / 84.4	**27.6 / 85.6**
cmn	17.2 / 79.5	18.5 / 80.5	13.4 / 74.0	20.9 / 82.8	17.7 / 81.3	20.8 / 83.2	**22.2 / 84.2**
dan	18.6 / 73.9	23.4 / 80.8	17.6 / 73.4	22.4 / 80.0	23.3 / 82.5	26.8 / 84.3	**29.1 / 86.0**
deu	27.6 / 83.8	25.8 / 83.4	19.3 / 74.9	26.9 / 83.8	25.2 / 83.9	28.4 / 85.5	**29.1 / 85.8**
ell	13.5 / 67.7	21.1 / 80.1	14.3 / 71.1	12.9 / 69.3	20.1 / 80.2	23.4 / 82.4	**25.7 / 84.4**
eng	34.2 / 86.6	30.5 / 84.3	31.8 / 85.3	31.9 / 85.7	31.9 / 87.1	34.2 / 87.9	**36.1 / 88.7**
fas	22.6 / 80.6	19.1 / 77.8	14.9 / 69.9	14.2 / 71.6	19.0 / 80.1	23.3 / 83.4	**25.3 / 84.7**
fin	16.4 / 75.5	21.9 / 82.7	13.9 / 69.5	17.8 / 78.5	21.2 / 82.9	24.5 / 84.9	**26.3 / 86.0**
fra	27.0 / 83.6	24.9 / 83.6	19.2 / 77.0	25.9 / 83.8	24.4 / 84.1	27.1 / 85.4	**28.7 / 86.3**
heb	12.9 / 62.6	18.7 / 73.2	15.6 / 67.6	7.4 / 56.6	15.4 / 72.7	20.4 / 77.1	**23.3 / 79.6**
hin	23.7 / 83.5	17.9 / 78.4	13.2 / 71.9	20.0 / 81.4	21.9 / 83.3	23.3 / 84.0	**25.8 / 85.5**
hrv	19.5 / 76.7	22.1 / 80.8	16.6 / 70.3	15.7 / 75.9	23.2 / 83.0	25.8 / 84.5	**27.9 / 85.7**
hun	14.8 / 70.6	21.2 / 79.9	15.3 / 68.9	11.0 / 65.0	19.6 / 79.9	23.8 / 83.4	**26.1 / 85.0**
ind	27.4 / 84.6	23.7 / 82.2	15.2 / 71.5	24.7 / 83.4	24.2 / 84.2	27.3 / 85.4	**28.7 / 86.4**
ita	25.6 / 84.4	22.7 / 83.1	17.1 / 75.8	23.3 / 83.4	22.5 / 83.8	25.6 / **85.6**	**26.1 / 85.6**
jpn	18.4 / 80.8	18.9 / 80.5	11.9 / 69.5	19.5 / 81.8	18.0 / 80.8	20.4 / 82.3	**22.0 / 84.0**
kaz	16.1 / 76.0	15.2 / 73.5	12.5 / 70.0	3.0 / 47.8	12.4 / 72.0	19.2 / 79.9	**21.4 / 82.1**
khm	11.1 / 71.1	0.0 / 28.8	8.9 / 65.6	1.1 / 43.3	4.8 / 61.1	11.6 / 71.5	**12.2 / 72.4**
kor	20.8 / 83.1	19.5 / 81.8	14.2 / 74.1	20.4 / 82.7	20.9 / 83.4	22.4 / 84.2	**24.4 / 85.6**
lao	14.0 / 73.4	6.3 / 59.9	10.6 / 67.3	10.2 / 66.9	8.3 / 67.8	14.7 / 74.7	**16.7 / 77.6**
msa	23.6 / 81.2	22.1 / 80.4	12.9 / 67.4	21.0 / 79.7	22.9 / 82.5	25.5 / **84.1**	**27.0** / 84.0
mya	6.1 / 65.1	0.1 / 28.0	**8.9 / 65.6**	0.9 / 43.2	1.4 / 49.9	6.2 / 60.5	4.6 / 57.0
nld	21.6 / 79.9	22.1 / 82.3	16.5 / 73.7	22.0 / 82.2	22.2 / 83.3	24.1 / 84.6	**26.2 / 85.7**
nob	22.6 / 78.8	22.3 / 80.4	16.5 / 72.3	23.0 / 81.1	24.1 / 83.6	27.1 / 85.0	**28.8 / 86.1**
pol	19.8 / 78.5	21.0 / 81.2	13.9 / 69.9	20.0 / 80.1	21.7 / 82.9	23.6 / 84.1	**25.5 / 85.2**
por	**30.9** / 85.6	26.2 / 83.5	16.3 / 71.8	27.3 / 83.8	26.5 / 85.0	29.4 / 86.2	30.5 / **86.9**
ron	23.9 / 79.8	24.8 / 81.9	18.9 / 73.1	22.5 / 79.8	24.8 / 83.3	27.6 / 84.9	**29.0 / 85.9**
rus	26.7 / 83.8	23.7 / 82.8	18.3 / 75.3	24.4 / 82.6	24.1 / 83.8	26.2 / 84.8	**27.9 / 85.5**
slk	22.6 / 79.9	23.3 / 82.0	17.9 / 73.2	17.5 / 76.6	23.4 / 83.1	26.4 / 84.7	**28.4 / 86.1**
slv	15.1 / 71.3	19.6 / 77.7	14.8 / 68.0	11.9 / 67.0	19.7 / 79.6	22.9 / 82.1	**25.6 / 84.6**
spa	26.1 / 85.3	22.8 / 83.4	15.9 / 75.2	23.0 / 83.3	23.1 / 84.6	25.3 / 85.7	**26.2 / 86.2**
swe	20.9 / 77.4	24.5 / 82.8	17.6 / 72.8	22.0 / 79.6	25.3 / 84.6	27.8 / 85.7	**29.7 / 86.8**
tam	17.3 / 78.9	12.9 / 71.4	12.6 / 69.8	2.5 / 50.4	13.0 / 74.0	17.9 / 79.4	**20.3 / 81.8**
tgl	22.4 / 79.1	21.8 / 79.4	11.5 / 60.2	17.9 / 74.6	18.6 / 78.1	25.2 / 82.7	**25.4 / 83.0**
tha	20.2 / 82.0	17.1 / 78.6	11.5 / 69.0	18.8 / 81.3	15.6 / 78.9	20.8 / 82.9	**22.4 / 83.9**
tur	25.6 / 83.7	23.4 / 83.0	15.8 / 72.1	23.2 / 82.5	24.0 / 84.1	25.7 / 85.2	**28.2 / 86.4**
urd	19.6 / 80.4	16.2 / 75.3	11.3 / 68.2	11.0 / 68.3	17.3 / 79.2	20.0 / 81.1	**23.2 / 83.9**
uzb	14.1 / 72.4	5.6 / 53.9	14.7 / 70.5	2.9 / 46.6	10.0 / 66.7	17.6 / 76.5	**21.6 / 81.1**
vie	19.7 / 80.1	18.7 / 78.9	14.1 / 72.6	20.0 / 81.3	18.0 / 80.5	21.4 / 82.7	**23.3 / 84.1**
yue	7.9 / 67.9	16.8 / 78.6	10.9 / 67.9	19.0 / 81.8	14.0 / 77.6	18.0 / 80.8	**19.8 / 82.1**
Avg.	20.3 / 78.3	19.4 / 76.5	15.3 / 71.5	17.3 / 74.1	19.7 / 79.6	23.2 / 82.5	**25.0 / 83.8**

Avg. is identical across Tables 6–7, as both average the same 1,980 direction pairs.

Table 7: spBLEU / COMET results on $44 \rightarrow X$ directions.

ISO	Gemini-3.5 -Flash-Lite	Whisper+ NLLB-3.3B	SeamlessM4T -V2-Large	Qwen3-Omni -30B-Instruct	MCAT -27B	MSRT-4B (ours)	MSRT-12B (ours)
ara	18.6 / 74.1	20.9 / 75.8	16.8 / 71.2	19.1 / 73.7	21.1 / 78.4	23.2 / 80.5	**25.3 / 81.7**
azj	15.1 / 77.8	14.1 / 75.2	11.9 / 70.1	11.3 / 70.8	12.3 / 77.0	16.7 / 81.2	**17.9 / 82.8**
ben	18.3 / 75.7	18.8 / 75.0	15.4 / 70.5	14.8 / 70.3	17.2 / 76.0	20.6 / 79.0	**22.0 / 80.1**
bul	25.0 / 81.1	25.5 / 79.9	19.3 / 73.4	21.5 / 75.8	25.1 / 82.7	29.3 / 85.4	**31.0 / 86.6**
cat	23.6 / 77.2	24.3 / 76.2	18.9 / 72.3	20.4 / 74.2	23.5 / 78.7	27.8 / 81.7	**29.7 / 83.0**
ces	21.0 / 81.7	20.4 / 79.4	15.9 / 71.8	17.8 / 77.5	19.8 / 83.3	22.7 / 85.8	**24.8 / 87.4**
cmn	21.1 / 78.3	14.3 / 71.2	11.3 / 65.4	21.3 / 78.4	21.5 / 80.2	25.2 / 83.3	**27.4 / 84.5**
dan	22.4 / 80.1	23.4 / 79.0	17.6 / 73.9	19.9 / 76.3	22.5 / 81.6	26.9 / 84.4	**28.7 / 85.7**
deu	22.2 / 76.5	22.7 / 75.0	15.3 / 64.8	21.3 / 74.8	21.9 / 78.1	25.0 / 80.8	**27.3 / 82.4**
ell	21.6 / 79.4	22.0 / 78.5	17.6 / 73.3	16.1 / 72.8	21.1 / 81.3	24.9 / 83.9	**26.6 / 85.2**
eng	24.9 / 79.2	31.7 / 80.5	31.1 / 83.3	28.1 / 79.7	27.5 / 80.6	34.0 / 84.1	**35.6 / 85.2**
fas	20.0 / 77.6	18.7 / 75.8	13.9 / 68.1	16.5 / 72.3	19.7 / 78.9	21.7 / 81.4	**23.5 / 82.9**
fin	20.2 / 82.1	16.6 / 78.2	13.1 / 72.7	14.6 / 76.2	18.4 / 83.6	20.8 / 85.4	**23.0 / 87.3**
fra	26.9 / 75.5	29.4 / 75.9	21.5 / 69.3	26.7 / 75.0	27.0 / 77.3	30.5 / 80.4	**32.6 / 81.7**
heb	22.4 / 77.3	21.3 / 75.4	15.9 / 68.9	11.4 / 64.6	22.1 / 79.4	24.8 / 81.8	**27.5 / 83.4**
hin	18.9 / 67.0	20.0 / 66.2	15.8 / 62.2	15.5 / 62.1	18.7 / 67.7	20.4 / 70.0	**22.8 / 71.4**
hrv	20.2 / 81.7	18.4 / 79.0	13.3 / 71.9	16.7 / 77.2	18.5 / 82.8	22.2 / 85.4	**23.9 / 86.9**
hun	20.6 / 79.7	18.4 / 75.7	13.8 / 67.9	15.9 / 74.4	17.8 / 79.7	21.4 / 82.4	**23.3 / 84.2**
ind	21.7 / 80.9	22.9 / 80.5	16.1 / 76.7	21.0 / 79.0	21.9 / 82.1	25.3 / 85.0	**27.0 / 86.1**
ita	19.9 / 77.7	20.1 / 77.1	14.6 / 71.4	19.2 / 76.5	19.9 / 79.8	22.0 / 82.5	**24.4 / 84.0**
jpn	18.2 / 81.8	9.3 / 76.1	7.3 / 71.9	17.2 / 81.0	18.2 / 83.1	19.9 / 85.4	**22.1 / 86.6**
kaz	19.7 / 81.3	16.8 / 78.9	13.8 / 72.8	13.7 / 73.4	16.0 / 81.2	21.0 / 84.6	**23.8 / 85.9**
khm	16.6 / 73.6	13.4 / 72.2	10.2 / 70.1	10.5 / 68.4	11.3 / 72.6	17.9 / 78.5	**18.6 / 79.2**
kor	16.0 / 79.9	13.9 / 77.5	10.2 / 71.8	15.2 / 78.7	15.5 / 80.7	16.8 / 83.2	**18.8 / 84.3**
lao	18.2 / 73.3	18.0 / 74.0	11.7 / 69.0	14.9 / 69.1	17.6 / 75.3	22.7 / 79.1	**24.7 / 79.9**
msa	19.8 / 78.8	20.7 / 78.5	13.1 / 73.8	15.5 / 76.1	19.0 / 79.8	23.2 / 83.0	**24.7 / 83.8**
mya	15.8 / 80.4	9.4 / 77.4	10.2 / 75.0	6.2 / 62.5	10.9 / 78.1	17.3 / 83.3	**19.4 / 84.6**
nld	19.4 / 77.5	20.0 / 76.4	14.6 / 69.3	18.0 / 74.7	19.1 / 79.5	22.4 / 82.0	**23.5 / 83.2**
nob	18.0 / 79.5	19.1 / 78.3	14.7 / 73.1	16.7 / 77.2	17.8 / 81.0	21.1 / 83.8	**22.6 / 85.1**
pol	19.0 / 80.4	17.9 / 78.3	13.9 / 70.8	16.5 / 76.8	17.8 / 82.5	20.5 / 84.7	**22.1 / 86.4**
por	24.1 / 78.2	25.8 / 78.0	20.4 / 74.6	23.7 / 76.6	25.0 / 79.9	28.8 / 82.7	**30.8 / 84.1**
ron	23.6 / 79.6	21.1 / 78.2	18.7 / 73.7	20.4 / 75.3	23.2 / 81.3	27.1 / 84.2	**29.2 / 85.5**
rus	21.8 / 80.0	22.3 / 78.8	16.0 / 71.9	20.5 / 76.3	21.7 / 82.1	24.6 / 84.6	**26.2 / 86.0**
slk	20.1 / 80.9	19.9 / 78.4	15.2 / 71.8	16.0 / 75.0	18.3 / 82.2	22.1 / 85.0	**23.7 / 86.6**
slv	19.8 / 80.6	17.7 / 77.6	13.8 / 72.1	15.3 / 75.7	17.5 / 81.7	21.5 / 84.3	**23.3 / 86.1**
spa	18.4 / 76.0	19.7 / 75.9	15.0 / 71.7	18.1 / 75.2	18.6 / 77.8	21.5 / 80.6	**22.7 / 81.9**
swe	22.9 / 79.7	23.9 / 78.4	18.5 / 74.1	20.5 / 76.1	23.2 / 81.3	27.0 / 84.1	**28.9 / 85.6**
tam	20.0 / 80.0	19.8 / 78.5	17.6 / 75.9	15.1 / 74.7	19.0 / 81.0	21.2 / 83.1	**23.7 / 84.2**
tgl	19.1 / 76.1	18.5 / 74.1	15.9 / 72.5	13.6 / 70.6	19.2 / 77.1	21.8 / 79.5	**22.9 / 79.9**
tha	27.6 / 78.1	20.2 / 74.0	20.1 / 70.3	26.7 / 78.1	26.2 / 79.8	29.9 / 82.5	**31.8 / 83.6**
tur	19.6 / 78.4	18.3 / 76.0	13.6 / 70.0	15.8 / 73.8	18.6 / 78.9	21.1 / 81.6	**23.1 / 83.4**
urd	16.6 / 72.3	17.2 / 70.8	13.8 / 67.2	12.9 / 65.9	14.7 / 72.6	17.7 / 75.6	**19.3 / 77.0**
uzb	19.6 / 81.5	17.5 / 79.3	14.8 / 75.5	8.7 / 67.7	17.0 / 81.8	22.2 / 84.3	**24.0 / 85.4**
vie	23.5 / 79.4	24.0 / 78.7	18.4 / 73.8	22.2 / 77.3	23.2 / 81.1	27.1 / 83.9	**28.4 / 84.8**
yue	11.7 / 75.5	6.2 / 68.5	7.9 / 67.6	16.2 / 78.1	18.7 / 81.4	20.8 / 83.9	**22.0 / 84.1**
Avg.	20.3 / 78.3	19.4 / 76.5	15.3 / 71.5	17.3 / 74.1	19.7 / 79.6	23.2 / 82.5	**25.0 / 83.8**

Avg. is identical across Tables 6–7, as both average the same 1,980 direction pairs.

Table 8: spBLEU / COMET results on English → 44 directions.

ISO	Gemini-3.5 -Flash-Lite	Whisper+ NLLB-3.3B	SeamlessM4T -V2-Large	Qwen3-Omni -30B-Instruct	MCAT -27B	MSRT-4B (ours)	MSRT-12B (ours)
ara	34.4 / 83.8	35.5 / 83.9	35.1 / 84.5	37.3 / 86.6	35.7 / 86.1	35.6 / 86.0	**38.4 / 86.8**
azj	22.0 / 85.8	20.7 / 84.0	20.7 / 83.8	17.9 / 83.0	17.8 / 84.8	21.9 / 87.1	**22.8 / 87.9**
ben	28.7 / 85.1	29.7 / 84.3	30.0 / 84.6	25.9 / 83.3	26.9 / 85.2	30.0 / 86.2	**31.1 / 86.6**
bul	43.7 / 88.9	41.4 / 87.7	43.5 / 88.8	41.6 / 89.1	42.2 / 90.0	45.4 / 90.3	**46.1 / 91.2**
cat	41.5 / 85.3	39.0 / 84.1	40.5 / 85.0	40.7 / 86.3	39.6 / 85.9	43.4 / 86.9	**44.8 / 87.6**
ces	37.2 / 89.1	33.3 / 87.3	34.8 / 88.0	34.5 / 89.0	33.8 / 90.1	34.9 / 90.4	**38.1 / 91.5**
cmn	33.4 / 86.4	19.5 / 75.3	23.0 / 79.7	37.0 / 88.4	33.0 / 87.2	34.5 / 87.8	**37.6 / 88.8**
dan	43.4 / 88.7	40.1 / 87.7	43.5 / 88.7	42.6 / 89.0	41.6 / 89.5	45.2 / 90.1	**47.1 / 90.8**
deu	41.8 / 86.2	38.5 / 84.4	39.2 / 84.9	**43.5** / 86.8	39.2 / 86.5	39.4 / 86.9	41.9 / **87.6**
ell	34.2 / 87.1	32.8 / 86.2	33.3 / 87.5	31.6 / 87.0	33.4 / 88.7	36.2 / 89.1	**38.0 / 90.0**
fas	33.2 / 86.5	29.1 / 83.7	31.6 / 84.6	30.6 / 85.6	32.3 / 86.9	32.1 / 87.1	**34.3 / 88.3**
fin	**35.4** / 90.9	27.3 / 87.9	29.0 / 88.5	27.1 / 88.6	30.6 / 91.3	31.6 / 91.5	33.7 / **92.4**
fra	48.4 / 85.8	47.1 / 84.7	48.2 / 85.3	**51.2** / 87.5	46.1 / 86.3	46.5 / 86.8	48.9 / **87.6**
heb	39.7 / 85.2	36.7 / 84.5	37.9 / 84.5	22.6 / 73.8	37.8 / 87.2	40.3 / 87.8	**43.1 / 88.7**
hin	32.0 / 79.0	33.9 / 78.1	33.8 / 78.1	30.2 / 77.6	30.5 / 78.9	32.1 / 79.3	**34.8 / 80.2**
hrv	**35.8** / 89.0	30.0 / 86.7	32.3 / 87.8	31.5 / 87.6	31.1 / 89.2	33.5 / 90.0	35.2 / **91.0**
hun	**33.7** / 88.1	29.0 / 85.6	31.0 / 86.0	28.4 / 87.0	27.8 / 87.6	30.8 / 88.4	32.3 / **89.4**
ind	39.9 / 89.8	38.8 / 88.6	38.2 / 89.0	**42.9** / 91.0	39.1 / 90.2	40.0 / 90.7	42.0 / **91.3**
ita	32.7 / 86.3	30.5 / 85.2	30.7 / 85.1	**34.6** / 87.3	31.1 / 86.8	31.8 / 87.3	33.5 / **88.1**
jpn	29.2 / 89.7	14.1 / 84.9	15.9 / 84.7	30.1 / 90.9	28.8 / 90.4	29.0 / 90.6	**31.0 / 91.2**
kaz	32.3 / 88.6	25.8 / 87.4	28.9 / 87.9	24.6 / 85.0	26.0 / 88.2	30.9 / 89.9	**33.9 / 90.4**
khm	23.5 / 81.4	16.0 / 74.9	16.7 / 79.9	15.3 / 77.5	16.2 / 79.7	**23.8** / 83.9	23.6 / **84.0**
kor	**27.0** / 88.4	21.7 / 86.7	21.5 / 85.1	26.9 / **89.7**	24.6 / 88.5	24.5 / 89.0	26.9 / 89.4
lao	27.5 / 81.4	23.3 / 79.3	24.8 / 81.4	25.6 / 80.3	27.1 / 83.1	30.6 / 84.5	**34.3 / 85.5**
msa	35.9 / 87.6	36.9 / 86.5	31.7 / 86.6	32.0 / 88.3	34.0 / 87.5	37.0 / 88.5	**38.4 / 89.0**
mya	21.9 / 87.1	14.2 / 79.9	17.2 / 85.7	9.9 / 71.7	15.2 / 85.0	22.5 / 88.3	**24.8 / 88.9**
nld	31.5 / 85.5	30.2 / 83.7	31.0 / 85.1	32.1 / 86.2	29.7 / 86.6	31.6 / 87.0	**32.5 / 87.9**
nob	33.8 / 87.9	32.3 / 86.2	33.7 / 86.8	33.3 / 88.4	31.5 / 88.7	34.5 / 89.2	**35.1 / 89.9**
pol	30.0 / 87.1	26.6 / 85.3	27.9 / 85.7	29.0 / 87.3	27.3 / 88.1	29.3 / 88.3	**30.2 / 89.7**
por	45.9 / 87.3	46.0 / 86.4	47.5 / 86.6	**50.1** / 88.6	45.3 / 87.9	46.9 / 88.2	49.0 / **89.3**
ron	42.0 / 88.5	35.2 / 86.9	39.9 / 87.8	41.0 / 88.8	38.9 / 89.1	41.8 / 89.8	**44.1 / 90.8**
rus	37.2 / 87.8	33.8 / 86.0	34.3 / 86.3	**38.6** / 88.9	34.8 / 88.8	36.2 / 88.8	37.7 / **89.9**
slk	**36.4** / 88.4	33.7 / 86.2	35.3 / 87.3	31.5 / 86.8	31.4 / 88.8	34.8 / 89.6	36.0 / **90.5**
slv	33.7 / 87.8	28.7 / 85.1	29.2 / 86.8	27.4 / 85.4	28.6 / 88.5	32.2 / 89.0	**33.8 / 90.4**
spa	29.6 / 84.2	29.3 / 83.7	28.1 / 83.2	30.7 / 85.4	29.7 / 85.2	30.7 / 85.5	**32.2 / 86.6**
swe	44.2 / 88.7	42.2 / 88.0	45.1 / 88.4	43.9 / 88.8	43.3 / 89.3	45.5 / 89.9	**47.1 / 90.4**
tam	31.5 / 87.6	30.0 / 84.4	32.4 / 87.3	25.6 / 85.6	29.3 / 88.3	31.1 / 88.5	**33.9 / 89.3**
tgl	31.8 / 82.9	29.7 / 82.0	30.2 / 82.1	24.0 / 80.0	31.1 / 83.3	32.3 / 83.9	**33.9 / 84.8**
tha	40.7 / 87.1	26.6 / 80.6	30.6 / 83.3	**41.1 / 88.9**	37.6 / 87.7	39.3 / 88.3	40.8 / **88.9**
tur	35.0 / 88.3	32.4 / 86.9	32.0 / 86.7	31.9 / 88.2	31.5 / 88.2	33.4 / 88.7	**35.7 / 89.5**
urd	26.4 / 81.2	26.0 / 78.6	25.9 / 79.4	23.0 / 78.3	22.8 / 80.9	26.1 / 82.0	**27.0 / 83.1**
uzb	31.0 / 88.2	27.1 / 87.7	28.6 / 87.7	16.4 / 79.2	25.6 / 88.2	31.9 / 89.6	**33.6 / 90.3**
vie	37.8 / 87.1	36.8 / 85.6	36.7 / 85.6	**40.2** / 88.6	35.9 / 87.8	38.4 / 88.3	40.1 / **89.0**
yue	16.4 / 81.7	9.3 / 76.1	15.7 / 79.8	29.4 / 88.6	27.7 / 88.0	27.9 / 88.3	**29.5 / 88.9**
Avg.	34.2 / 86.6	30.5 / 84.3	31.8 / 85.3	31.9 / 85.7	31.9 / 87.1	34.2 / 87.9	**36.1 / 88.7**