LCMA-SRT: Language-Conditional Mixture-of-Experts Adapters for Joint Multilingual Speech Recognition and Translation

Anonymous ACL submission

Abstract

Neural transducers provide an alignment-free framework for joint automatic speech recognition (ASR) and speech translation (ST). Hierarchical transducer architectures further improve multilingual speech-to-text modeling by stacking a translation-focused encoder on top of an ASR encoder to better handle reordering. However, scaling hierarchical transducers to multilingual many-to-many settings remains challenging: fully shared models often suffer from negative transfer and unstable target-language generation, while training separate models per direction is computationally prohibitive. We propose LCMA-SRT (Language-Conditional Mixture-of-Experts Adapters for Speech Recognition and Translation), which augments a hierarchical transducer with language-conditional Mixture-of-Experts (MoE) adapters. A source-conditioned MoE adapter (SRC-MoE) routes using the source-language embedding to improve acoustic–phonetic modeling and reduce cross-language interference for ASR. A target-conditioned MoE adapter (TGT-MoE) routes using the desired target language to guide reordering and lexical selection and to mitigate cross-target interference in many-to-many ST. Experiments on Europarl-ST (9 languages, 72 directions) show that LCMA-SRT improves both ASR and ST within a single joint model, reducing average WER and increasing BLEU and COMET over strong hierarchical transducer baselines. We release our codes and models at https://anonymous.4open.science/r/LCMA-SRT.

1 Introduction

Speech translation (ST) enables spoken content to be recognized and translated across languages, supporting cross-lingual communication in settings such as multilingual meetings, education, healthcare, and customer service (Köksal and Yürük, 2020; Al Shamsi et al., 2020). Conventional ST systems are typically cascades: an ASR module first transcribes speech into text, which is then translated by an MT system (Matusov et al., 2005; Bertoldi and Federico, 2005). Despite the simplicity of this modular design, cascades suffer from error propagation, duplicated computation, and increased system complexity when ASR and MT are optimized separately (Sperber et al., 2017; Rabatin et al., 2024).

End-to-end speech translation has therefore gained momentum as a more streamlined alternative, aiming to map speech directly to target text while sharing representations across ASR and translation (Berard et al., 2016, 2018). Most progress has been driven by attention-based encoder–decoder (AED) models (Gaido et al., 2020; Hussein et al., 2023), but they can be sensitive to segmentation and complicate joint ASR and ST optimization (Anastasopoulos et al., 2022; Sinclair et al., 2014). In contrast, alignment-free objectives such as CTC and neural transducers provide a frame-synchronous modeling paradigm that has been highly effective for speech recognition (Graves et al., 2006; Graves, 2012), and can naturally combine recognition and translation within one architecture (Chiu et al., 2019).

A key difficulty is that ST is typically less monotonic than ASR and often requires substantial reordering (Yan et al., 2023). As a result, training ASR and ST in a single transducer pipeline with fully shared representations can degrade translation quality, especially for language pairs with strong word-order divergence (Wang et al., 2023; Tang et al., 2023).

Hierarchical transducer architectures mitigate this mismatch by explicitly separating recognition and translation within one model. A recent work (Hussein et al., 2025) first produces speech-aligned, approximately monotonic representations via an ASR encoder and then stacks a translation-specific encoder to better model non-monotonic re-

ordering, improving transducer-based ST (Dalmia et al., 2021).

Scaling such hierarchical transducers to multilingual many-to-many ASR and ST remains challenging. Fully shared multilingual models often suffer from negative transfer under language imbalance and cross-language interference, leading to unstable direction control and language mixing on the translation side (Xue et al., 2024), while training separate direction-specific models is computationally impractical. Meanwhile, large unified systems (e.g., Whisper (Radford et al., 2023), HuBERT (Hsu et al., 2021), Whistle (Yusuyin et al., 2024), Seamless (Barrault et al., 2023), Canary (Sekoyan et al., 2025)) show strong multilingual transfer via scaling, but often at high model/compute cost. TTA (Liu et al., 2025) demonstrates that ASR and ST multi-task training with speech and text alignment and language tokens improves cross-lingual transfer; however, such control signals alone may be insufficient to fully prevent cross-target interference in a fully shared many-to-many transducer, motivating language-conditional specialization.

To reconcile scalability with accuracy, we argue that a unified many-to-many model should preserve the benefits of hierarchical transduction while introducing language-conditional specialization, rather than relying on fully shared parameters everywhere. Mixture-of-experts (MoE) offers conditional capacity to capture both shared and language-specific structure (Shazeer et al., 2017; Fedus et al., 2022), and lightweight MoE adapters have been shown to improve multilingual ASR robustness under data imbalance with a small parameter overhead (Li et al., 2025; Mu et al., 2024).

In this work, we propose LCMA-SRT, a unified multilingual hierarchical neural transducer framework that combines a hierarchical transducer encoder (Hussein et al., 2025) with language-conditional MoE adapters to scale joint ASR and ST to many-to-many settings. LCMA-SRT targets the central difficulty of many-to-many multilingual transduction: fully shared models often suffer from negative transfer on the recognition side and cross-target interference on the translation side, while direction-specific training is computationally prohibitive. Our objectives are three-fold: (1) to extend hierarchical transducer modeling from many-to-one to a single many-to-many model for joint ASR and ST, (2) to improve multilingual ASR robustness under heterogeneous acoustic–phonetic conditions without sacrificing cross-lingual sharing, and (3) to strengthen target-language control for multilingual ST by reducing cross-target interference and language mixing within a unified model. To this end, we introduce two complementary language-conditional adapter mechanisms: a source-conditioned MoE adapter (SRC-MoE) and a target-conditioned MoE adapter (TGT-MoE), integrated into the hierarchical transducer backbone. We evaluate LCMA-SRT on Europarl-ST (Iranzo-Sánchez et al., 2020) and perform controlled ablations to quantify the impact of each design component in many-to-many joint ASR and ST.

2 Proposed Approach

Given input speech features $X = (x_1, \ldots, x_T) \in \mathbb{R}^{T \times F}$, we consider two token sequences: a source-language transcript $\mathbf{y}^{(s)} = (y_1^s, \ldots, y_{U_s}^s) \in \mathcal{V}_{\mathrm{U_s}}^{(s)}$ (ASR) and a target-language translation $\mathbf{y}^{(t)} = (y_1^t, \ldots, y_{U_t}^t) \in \mathcal{V}_{\mathrm{U_t}}^{(t)}$ (ST) (Graves, 2012). For each task $k \in \{s, t\}$, a neural transducer models the conditional probability $P(\mathbf{y}^{(k)} \mid X)$ by marginalizing over all monotonic alignments $\mathbf{a} \in \bar{\mathcal{V}}_{\mathrm{T+U_k}}^{(k)}$:

$$P(\mathbf{y}^{(k)} \mid X) = \sum_{\mathbf{a} \in \mathcal{B}^{-1}(\mathbf{y}^{(k)})} P(\mathbf{a} \mid X) \quad (1)$$

$$\bar{\mathcal{V}}^{(k)} = \mathcal{V}^{(k)} \cup \{\phi\} \quad (2)$$

where ϕ denotes the blank symbol and $\mathcal{B}$ deterministically removes blanks to map an alignment to its corresponding label sequence. The transducer parameterization uses an encoder, a predictor, and a joiner to produce a posterior distribution over $\bar{\mathcal{V}}^{(k)}$ at each lattice state. To better handle the monotonicity mismatch between recognition and translation, a hierarchical transducer employs a stacked encoder hierarchy (Hussein et al., 2025). An ASR encoder first produces speech-aligned representations, which are then transformed by a translation-specific encoder:

$$\mathbf{F}^{(s)} = \mathrm{Enc}_{\mathrm{asr}}(X) \quad (3)$$

$$\mathbf{F}^{(t)} = \mathrm{Enc}_{\mathrm{st}}\left(\mathbf{F}^{(s)}\right) \quad (4)$$

Task-specific transducer heads are applied on top of $\mathbf{F}^{(s)}$ (ASR) and $\mathbf{F}^{(t)}$ (ST), respectively. We train the model with a multitask transducer objective:

$$\mathcal{L}_{\mathrm{nt}} = \alpha_{\mathrm{asr}} \mathcal{L}_{\mathrm{nt}}^{(s)} + \alpha_{\mathrm{st}} \mathcal{L}_{\mathrm{nt}}^{(t)} \quad (5)$$

$$\mathcal{L}_{\mathrm{nt}}^{(k)} = -\log P(\mathbf{y}^{(k)} \mid X) \quad (6)$$

where α_{asr} and α_{st} control the contribution of each task.

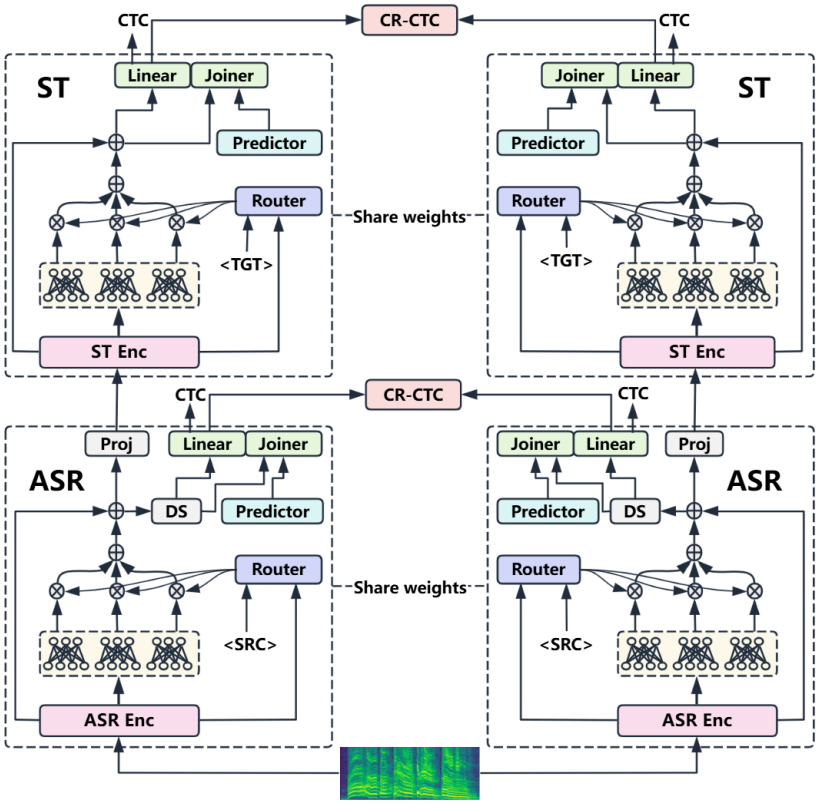

Figure 1: Proposed language-conditional mixture-of-experts adapters for unified multilingual speech recognition and translation (LCMA-SRT).

LCMA-SRT Overview: To improve multilingual ASR accuracy and extend hierarchical transducer modeling from many-to-one settings to a unified many-to-many setting, we propose LCMA-SRT. As illustrated in Figure 1, LCMA-SRT augments the hierarchical transducer backbone (Hussein et al., 2025) with two lightweight language-conditional MoE adapters inserted once at the encoder outputs: a SRC-MoE after the ASR encoder and TGT-MoE after the ST encoder.

2.1 ASR Encoder and SRC-MoE

As shown in Figure 1, the ASR branch takes the speech-aligned encoder representations $\mathbf{F}^{(s)}$ (defined in Eq. 3) and applies a SRC-MoE to mitigate cross-language interference in multilingual recognition. Let $\mathbf{F}^{(s)} = \{\mathbf{f}_t^{(s)}\}_{t=1}^{T}$ and let $\mathbf{e}_s = \mathrm{Embed}_{\mathrm{src}}(\ell_s) \in \mathbb{R}^{d_\ell}$ denote a learned embedding of the source language identity (corresponding to <SRC> in Figure 1). SRC-MoE consists of a router $g(\cdot)$ and E lightweight experts $\{f_i(\cdot)\}_{i=1}^{E}$ (Shazeer et al., 2017; Fedus et al., 2022). For notational clarity, we describe the adapter in a generic form using an input hidden vector $\mathbf{h}_t$; in the ASR branch we set $\mathbf{h}_t = \mathbf{f}_t^{(s)}$. The router conditions on the hidden state and the source-language embedding to produce mixture weights:

$$\pi_t = g([\mathbf{h}_t; \mathbf{e}_s]) \in \mathbb{R}^E \tag{7}$$

$$\mathbf{w}_t = \mathrm{softmax}(\pi_t) \in \mathbb{R}^E \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation. Each expert transforms $\mathbf{h}_t$, and the adapter output is computed as a weighted mixture with a residual connection:

$$\mathbf{m}_t = \sum_{i=1}^{E} w_{t,i}\, f_i(\mathbf{h}_t) \tag{9}$$

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t + \mathbf{m}_t \tag{10}$$

Stacking $\tilde{\mathbf{h}}_t$ over time yields the adapted sequence $\tilde{\mathbf{F}}^{(s)} = \{\tilde{\mathbf{f}}_t^{(s)}\}_{t=1}^{T}$, where $\tilde{\mathbf{f}}_t^{(s)}$ corresponds to $\tilde{\mathbf{h}}_t$ under the substitution $\mathbf{h}_t = \mathbf{f}_t^{(s)}$. By conditioning routing on the source language, SRC-MoE allocates capacity to language-specific acoustic–phonetic variation while retaining sharing in the backbone, thereby improving multilingual ASR robustness.

Following the SRC-MoE adapter, we apply a 2× downsampling module on the ASR branch before

the ASR transducer head:

$$\hat{\mathbf{F}}^{(s)} = \mathrm{DS}\Big(\tilde{\mathbf{F}}^{(s)}\Big) \in \mathbb{R}^{T/2\times d} \tag{11}$$

The ASR predictor–joiner head then defines $P(\mathbf{y}^{(s)} \mid X)$ using $\hat{\mathbf{F}}^{(s)}$ as encoder inputs. Meanwhile, $\tilde{\mathbf{F}}^{(s)}$ is forwarded to the ST branch (via the projection module in Figure 1), enabling improvements in source-side modeling to benefit downstream translation.

2.2 ST Encoder and TGT-MoE

We adopt a two-stage training recipe: we first pretrain the ASR branch, and then perform a second-stage joint training of ASR and ST on top of the pretrained model. During the second stage, the ST encoder consumes the refined ASR representations $\tilde{\mathbf{F}}^{(s)}$ produced by SRC-MoE (Section 2.1), which implicitly encode source-language characteristics learned from multilingual ASR training. These source-aware intermediate representations facilitate learning translation while maintaining effective cross-lingual sharing.

As shown in Figure 1, we first apply a lightweight projection to match the input dimension expected by the ST encoder:

$$\mathbf{G}^{(s)} = \mathrm{Proj}\Big(\tilde{\mathbf{F}}^{(s)}\Big) \tag{12}$$

The ST encoder then produces translation-oriented representations:

$$\mathbf{F}^{(t)} = \mathrm{Enc}_{\mathrm{st}}\Big(\mathbf{G}^{(s)}\Big) \tag{13}$$

where $\mathbf{F}^{(t)} = \{\mathbf{f}_t^{(t)}\}_{t=1}^{T}$.

We apply a TGT-MoE once at the ST encoder output. Let $\mathbf{e}_t = \mathrm{Embed}_{\mathrm{tgt}}(\ell_t) \in \mathbb{R}^{d_\ell}$ denote the learned embedding of the desired target language (corresponding to <TGT> in Figure 1). TGT-MoE follows Eqs. 7–10, where the hidden input is set to $\mathbf{h}_t = \mathbf{f}_t^{(t)}$ and the conditioning embedding is $\mathbf{e} = \mathbf{e}_t$. The adapted ST representations are:

$$\tilde{\mathbf{F}}^{(t)} = \mathrm{Adapter}_{\mathrm{tgt}}\Big(\mathbf{F}^{(t)}; \mathbf{e}_t\Big) \tag{14}$$

Conditioning routing on the target language guides reordering and lexical selection toward the desired output language, reducing cross-target interference and stabilizing target-language generation in many-to-many ST. The ST predictor–joiner head defines the translation distribution using $\tilde{\mathbf{F}}^{(t)}$ as encoder inputs, yielding $P(\mathbf{y}^{(t)} \mid X, \ell_t)$.

2.3 CR-CTC Self-Distillation

Following prior work on hierarchical transducer training for joint ASR and ST (Hussein et al., 2025), we adopt consistency-regularized CTC (CR-CTC) as an auxiliary self-distillation signal (Yao et al., 2024). CR-CTC constructs two stochastic views of the same utterance under shared parameters and encourages their CTC posteriors to be consistent.

We attach lightweight CTC heads to both branches. For ASR, the auxiliary losses are computed on the downsampled representations $\hat{\mathbf{F}}^{(s)}$ (Eq. 11); for ST, they are computed on the TGT-MoE output $\tilde{\mathbf{F}}^{(t)}$ (Eq. 14). We define the auxiliary loss for each task as a weighted sum of CTC and CR terms:

$$\mathcal{L}_{\mathrm{aux}}^{(s)} = \lambda_{\mathrm{ctc}}^{(s)}\mathcal{L}_{\mathrm{ctc}}^{(s)} + \lambda_{\mathrm{cr}}^{(s)}\mathcal{L}_{\mathrm{cr}}^{(s)} \tag{15}$$

$$\mathcal{L}_{\mathrm{aux}}^{(t)} = \lambda_{\mathrm{ctc}}^{(t)}\mathcal{L}_{\mathrm{ctc}}^{(t)} + \lambda_{\mathrm{cr}}^{(t)}\mathcal{L}_{\mathrm{cr}}^{(t)} \tag{16}$$

Here $\mathcal{L}_{\mathrm{ctc}}^{(k)}$ is the CTC negative log-likelihood, and $\mathcal{L}_{\mathrm{cr}}^{(k)}$ enforces view-consistent CTC posteriors via a symmetric KL divergence over valid frames. The coefficients $\lambda_{\mathrm{ctc}}^{(s)}, \lambda_{\mathrm{cr}}^{(s)}$ and $\lambda_{\mathrm{ctc}}^{(t)}, \lambda_{\mathrm{cr}}^{(t)}$ weight the CTC and CR-CTC terms for ASR and ST, respectively.

2.4 MoE Entropy Regularization

MoE routing may collapse to a small subset of experts. To encourage balanced expert utilization, we regularize the router to have high entropy (Shazeer et al., 2017). Let $\mathbf{w} \in \mathbb{R}^{T\times B\times E}$ denote the routing weights returned by an MoE module, where $\mathbf{w}_{t,b}$ is a probability simplex over E experts. The per-frame entropy is

$$H(\mathbf{w}_{t,b}) = -\sum_{i=1}^{E} w_{t,b,i} \log w_{t,b,i} \tag{17}$$

With an optional padding mask $\mathbf{M} \in \{0,1\}^{T\times B}$ ($M_{t,b} = 1$ indicates padded frames), we compute the mean entropy over valid frames as

$$\bar{H}(\mathbf{w}) = \frac{\sum_{t,b}(1 - M_{t,b})\, H(\mathbf{w}_{t,b})}{\sum_{t,b}(1 - M_{t,b})} \tag{18}$$

We apply entropy regularization to both SRC-MoE and TGT-MoE, and average the two weighted terms:

$$\mathcal{L}_{\mathrm{ent}} = -\frac{1}{2}\sum_{k\in\{\mathrm{s,t}\}} \lambda_{\mathrm{ent}}^{(k)} \cdot \bar{H}\Big(\mathbf{w}^{(k)}\Big) \tag{19}$$

Here $\lambda_{\text{ent}}^{(s)}$ and $\lambda_{\text{ent}}^{(t)}$ control the strength of entropy regularization for SRC-MoE and TGT-MoE, respectively.

Finally, we optimize the overall multitask transducer loss together with the auxiliary CR-CTC losses and the MoE entropy regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{nt}} + \alpha_{\text{asr}}\mathcal{L}_{\text{aux}}^{(s)} + \alpha_{\text{st}}\mathcal{L}_{\text{aux}}^{(t)} + \mathcal{L}_{\text{ent}} \quad (20)$$

2.5 Decoding for ASR and ST

We decode both ASR and ST with a modified transducer beam search using their task-specific predictor–joiner heads (Graves, 2012; Jain et al., 2019). During inference, we condition the model on the source and target language identities (i.e., the <SRC> and <TGT> signals in Figure 1), so that SRC-MoE and TGT-MoE routing remains active in the ASR and ST branches.

For ST, we enforce the translation direction with a target-language prefix token $\tau(\ell_t)$ (e.g., <2xx>) (Johnson et al., 2016) for all systems in this work. ST training labels are prepended with $\tau(\ell_t)$, and at inference we force the first emitted token to be the same prefix. For the direction-specific HENT-SRT-M2O×9 baseline, $\tau(\ell_t)$ is fixed within each model (one target language per model). The prefix token is removed for evaluation. Following prior transducer-based ST practice, we further apply a blank penalty during ST decoding to discourage excessive blank emissions (Hussein et al., 2025). However, a prefix-only constraint does not prevent target-language drift in fully shared models (Sec. 3.2).

2.6 Computation Efficiency

LCMA-SRT is designed to improve many-to-many multilingual ASR and ST without sacrificing the compactness of hierarchical transducer modeling: instead of training and deploying separate direction-specific systems, it supports all translation directions within a single shared backbone, avoiding costs that grow with the number of language pairs. The extra computation is confined to two lightweight language-conditional MoE adapters inserted once after Enc_{asr} and once after Enc_{st} (Figure 1); each adapter comprises a small router and E shallow experts (two-layer MLPs) with a residual connection, so the dominant cost still lies in the shared encoders and the adapter overhead scales mainly with E, not with the number of translation directions. In our implementation, both branches use the Zipformer encoder and we train the transducer with the pruned transducer loss (Yao et al., 2023; Kuang et al., 2022), preserving the efficiency of the hierarchical transducer backbone while enabling conditional capacity allocation via SRC-MoE and TGT-MoE. [1]

3 Experiments

3.1 Experimental setup

We conduct experiments on Europarl-ST (Iranzo-Sánchez et al., 2020), a multilingual speech translation benchmark with paired speech, source-language transcripts, and target-language translations. We consider nine languages (en, de, es, fr, it, nl, pl, pt, ro) and cover all ordered cross-lingual pairs, yielding 72 translation directions. We follow the official train/dev/test splits and report test-set results for both ASR and ST.

Data processing: Our experiments utilize the icefall framework, with the Lhotse toolkit (Żelasko et al., 2021) for speech data preparation and dataset construction. ASR pretraining uses 276 hours of unique audio. For joint ASR and ST training, we reuse the same segments and pair each utterance with its available translations into the remaining target languages, yielding 1,642 pair-hours. All audio recordings are resampled to 16 kHz. We extract 80-dimensional log-Mel filterbank features using a 25 ms window with a 10 ms frame shift. Additionally, we apply on-the-fly SpecAugment (Park et al., 2019), incorporating time warping (maximum factor of 80), frequency masking (two regions, max width of 27 bins), and time masking (ten regions, max width of 100 frames). For CR-CTC self-distillation, we strengthen the augmentation of the second view by increasing both the number of time-masked regions and the maximum masking fraction by a factor of 2.5, following prior practice (Yao et al., 2024). For text, we apply Whisper-style normalization rules (Radford et al., 2023) prior to subword tokenization. For subword tokenization, we train SentencePiece BPE models (Kudo and Richardson, 2018) with a vocabulary size of 1k for ASR and 1k for ST. On the ASR side, we use a single multilingual BPE model shared across all source-language transcripts. On the ST side, we include a target-language tag token (e.g., <2xx>) in the ST vocabulary and prepend the corresponding tag to every ST training label sequence to indicate

[1] https://github.com/k2-fsa/icefall

the desired output language. For all multilingual many-to-many systems, we use a shared 1k ST BPE model. For the direction-specific baseline trained as nine separate systems (one per target language), we instead train an independent 1k ST BPE model for each target language, while keeping all other preprocessing settings identical.

Models: We reproduce a multitask hierarchical neural transducer for joint ASR and ST, employing the Zipformer encoder architecture and task-specific transducer heads. The encoder is partitioned into an ASR encoder stacked below an ST encoder, and we keep the Zipformer configuration identical to HENT-SRT for fair comparison. For Stage 1 ASR pretraining, we train three ASR-only variants: CR-CTC, CR-CTC+S-Bias (with a source-identity bias added to the encoder outputs), and CR-CTC+SRC-MoE (with the source-conditioned MoE adapter applied after the ASR encoder). The first two models have 30M parameters; adding SRC-MoE increases the parameter count to 36M. For Stage 2 joint ASR and ST training, we compare (i) direction-specific many-to-one training as nine separate models HENT-SRT-M2O$\times$9 (each 61M), (ii) HENT-SRT-M2M, a fully shared many-to-many extension of the original many-to-one HENT-SRT (61M; Sec. 2.5), and (iii) LCMA-SRT and its ablations, which examine the SRC-MoE and dissect the TGT-MoE into an unconditioned MoE and a simple target-identity bias. Unless otherwise specified, we use E=8 experts for SRC-MoE and E=16 experts for TGT-MoE, resulting in model sizes ranging from 66M to 77M parameters. Both tasks use a stateless transducer predictor implemented as a single Conv1D layer with kernel size 2, together with a standard joiner.

Our training uses the ScaledAdam optimizer (Yao et al., 2023) with a learning rate of 0.02 and a 5k-step warmup. We train in two stages on 4 NVIDIA A800 GPUs, each for 50 epochs, using duration-based batching with max-duration=900 s in Stage 1 and 450 s in Stage 2. Stage 2 is initialized from the best checkpoint of Stage 1. The multitask weights α_{asr} and α_{st} in Eq. (5) are both set to 1. For the auxiliary CR-CTC objective (Sec. 2.3), we set $\lambda_{\text{ctc}}^{(s)} = \lambda_{\text{ctc}}^{(t)} = 0.1$ and $\lambda_{\text{cr}}^{(s)} = \lambda_{\text{cr}}^{(t)} = 0.05$ for the CTC and CR terms, respectively. When MoE adapters are enabled, we apply router entropy regularization with $\lambda_{\text{ent}}^{(\text{s})} = \lambda_{\text{ent}}^{(\text{t})} = 0.015$. All decoding results are obtained with beam search using a beam size of 20. For ST, we apply a blank penalty of 2.

Evaluation: We evaluate ASR with word error rate (WER). For speech translation, we report BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and sentence-level target-language mismatch rate (LMR) using an off-the-shelf language identification model (Joulin et al., 2016): a hypothesis is matched only if it is classified as the specified target language with confidence ≥ 0.7. To ensure consistent comparisons, we normalize ASR transcripts and ST outputs by removing punctuation and evaluating case-insensitively.

3.2 Main Results

We evaluate our proposed framework on Europarl-ST following the two-stage training recipe. In this section, we analyze the effectiveness of the proposed adapters in ASR pretraining, followed by a comprehensive comparison against strong baselines in the joint ASR and ST setting.

Multilingual ASR Pretraining: Table 1 reports pretraining results. Relative to the CR-CTC baseline (22.35% Avg. WER), an unconditioned MoE yields a modest gain (22.06%), while adding a source-identity bias reduces WER to 21.08%. The proposed SRC-MoE achieves the best performance (20.88%), obtaining the lowest WER on 8/9 languages, suggesting that source-conditioned routing learns more robust speech-aligned representations for subsequent training.

Many-to-Many Joint Training: We present the many-to-many joint training results in Table 2 and Table 3. We evaluate LCMA-SRT against the direction-specific baseline HENT-SRT-M2O$\times$9. Notably, LCMA-SRT significantly outperforms this strong baseline, yielding an average improvement of +5.2 BLEU (15.3 $\rightarrow$ 20.5) and +0.076 COMET (0.575 $\rightarrow$ 0.651). We observe consistent gains across diverse data conditions: high-resource directions achieve substantial absolute improvements (e.g., Spanish: 19.1 $\rightarrow$ 25.8 BLEU), while lower-resource directions see remarkable relative gains (e.g., Polish: 7.2 $\rightarrow$ 10.7 BLEU, +48% relative improvement). This suggests that LCMA-SRT enables effective positive transfer from high-resource to low-resource languages within the shared MoE framework. Comparison with the fully shared baseline HENT-SRT-M2M further highlights the advantage of explicit target conditioning. Despite forced target-prefix decoding (Sec. 2.5), the fully shared HENT-SRT-M2M collapses under 72 competing directions due to severe

Model	WER (%) ↓									
	de	en	es	fr	it	nl	pl	pt	ro	Avg
CR-CTC	24.57	18.59	20.76	19.24	17.33	36.75	25.28	19.82	18.77	22.35
+ MoE	24.39	18.41	20.16	18.61	17.28	36.83	24.36	19.70	18.79	22.06
+ S-Bias	23.89	17.60	19.58	17.41	16.73	**34.72**	23.63	18.21	17.97	21.08
+ SRC-MoE	**23.34**	**17.45**	**19.41**	**17.34**	**16.27**	35.20	**23.28**	**18.16**	**17.48**	**20.88**

Table 1: Multilingual ASR results on Europarl-ST. WER is reported per source language, and Avg denotes the overall average. We report the CR-CTC baseline and its variants with an unconditioned MoE adapter (+MoE), a source-identity bias (+S-Bias), and the proposed source-conditioned MoE adapter inserted after the ASR encoder (+SRC-MoE).

Model	WER (%)↓	Average BLEU ↑									
		de	en	es	fr	it	nl	pl	pt	ro	Avg
HENT-SRT-M2O×9	23.28	10.7	21.2	19.1	18.2	14.2	16.5	7.2	18.4	12.1	15.3
HENT-SRT-M2M	16.65	2.6	12.8	5.5	4.0	1.8	3.5	1.2	4.9	2.5	4.3
LCMA-SRT	**15.71**	**15.2**	**25.9**	**25.8**	**24.7**	**20.0**	**20.5**	**10.7**	**23.9**	**17.6**	**20.5**
TGT-MoE→MoE	16.42	2.3	14.7	4.7	3.3	1.7	2.7	1.1	4.5	2.0	4.1
TGT-MoE→T-Bias	15.84	13.1	22.7	23.5	22.3	17.7	18.1	8.3	21.8	14.5	18.0
w/o TGT-MoE	16.48	2.0	12.8	5.9	3.9	1.6	3.0	1.3	5.0	2.2	4.2
w/o SRC-MoE	16.11	14.5	24.9	25.0	24.6	19.6	20.0	10.5	23.7	17.5	20.0

Table 2: Joint ASR and ST results on Europarl-ST. WER is averaged over all 72 translation directions. BLEU is averaged over directions grouped by their target language, and Avg denotes the overall average across all directions. We compare HENT-SRT-M2O×9 and HENT-SRT-M2M against LCMA-SRT and ablations that replace TGT-MoE with an unconditioned MoE (TGT-MoE→MoE) or a target-identity bias (TGT-MoE→T-Bias), or remove TGT-MoE / SRC-MoE (w/o TGT-MoE, w/o SRC-MoE).

Model	LMR (%)↓	Average COMET ↑									
		de	en	es	fr	it	nl	pl	pt	ro	Avg
HENT-SRT-M2O×9	**0.65**	0.507	0.656	0.587	0.542	0.565	0.558	0.550	0.609	0.598	0.575
HENT-SRT-M2M	84.95	0.380	0.543	0.478	0.427	0.435	0.401	0.385	0.471	0.406	0.436
LCMA-SRT	<u>0.75</u>	**0.574**	**0.715**	**0.682**	**0.627**	**0.656**	**0.613**	**0.616**	**0.693**	**0.678**	**0.651**
TGT-MoE→MoE	85.23	0.380	0.559	0.476	0.426	0.438	0.395	0.386	0.472	0.408	0.438
TGT-MoE→T-Bias	0.78	0.529	0.675	0.642	0.583	0.612	0.563	0.562	0.651	0.621	0.604
w/o TGT-MoE	85.19	0.376	0.545	0.480	0.427	0.434	0.398	0.387	0.473	0.407	0.436
w/o SRC-MoE	0.81	0.568	0.708	0.671	0.621	0.646	0.606	0.605	0.685	0.675	0.643

Table 3: Average COMET and LMR scores for the same models and ablations as Table 2.

cross-direction interference and target-language drift (LMR 84.95%), resulting in poor translation quality (4.3 BLEU). In contrast, LCMA-SRT maintains a low LMR (0.75%), comparable to the direction-specific HENT-SRT-M2O×9 (0.65%), while substantially improving translation quality (+5.2 BLEU, +0.076 COMET) and reducing WER to 15.71%. These findings demonstrate that LCMA-SRT effectively scales hierarchical transducer modeling to the many-to-many setting within a single unified model, achieving performance that outperforms the strong hierarchical transducer baselines reproduced in this work on both tasks simultaneously. Direction-wise ASR WER results are reported in Appendix A Table 4; direction-wise ST BLEU/COMET/LMR results are reported in Appendix A Tables 5, 6, and 7.

3.3 Ablation Analysis

To assess the contribution of each component in LCMA-SRT, we conduct ablations in Table 2 and Table 3. We first examine the critical role of Target-Conditioned MoE (TGT-MoE). Removing TGT-MoE (w/o TGT-MoE) leads to a collapse in translation quality (4.2 BLEU) and an extremely high LMR (85.19%), showing that a monolithic ST encoder fails to preserve target-language fidelity under 72 competing directions. Replacing condi-

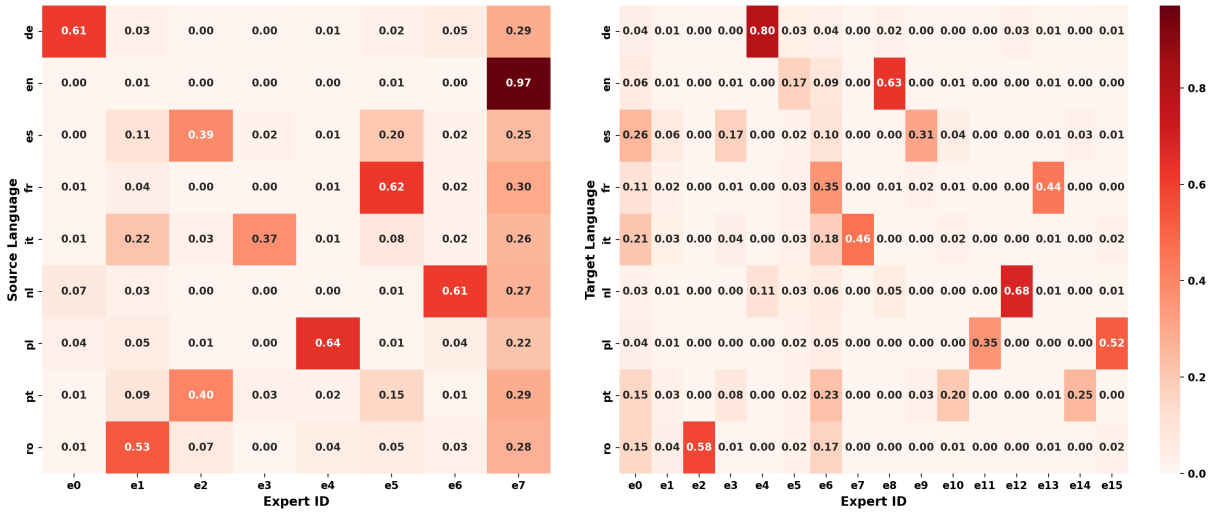

Figure 2: Heatmaps of language-aggregated expert routing weights in the language-conditional MoE adapters. Left: SRC-MoE on the ASR side, where each row is a source language and each column is an expert (e0–e7). Right: TGT-MoE on the ST side, where each row is a target language and each column is an expert (e0–e15).

tional routing with an unconditioned MoE (TGT-MoE→MoE) yields similarly high LMR (85.23%) and poor BLEU, demonstrating that merely increasing parameters is insufficient—explicit target conditioning is essential to disentangle languages. Interestingly, a target-identity bias (TGT-MoE→T-Bias) restores low LMR (0.78%), indicating that directional cues help avoid language leakage; however, it still lags behind full TGT-MoE in BLEU/COMET, suggesting that expert-based non-linear capacity is necessary beyond simply selecting the right language. For the ASR encoder, removing the SRC-MoE (w/o SRC-MoE) leads to a degradation in ASR WER (15.71% → 16.11%) and a consequent drop in ST BLEU (20.5 → 20.0), validating that mitigating source-side acoustic interference via conditional adapters yields better speech-aligned representations that benefit downstream translation.

3.4 Analysis of MoE Routing Behavior

Figure 2 visualizes the language-aggregated routing distributions learned by our language-conditional MoE adapters under entropy regularization. On the ASR side (SRC-MoE, left), each expert is biased toward certain source languages (e.g., e_4 is strongly activated by pl), while several experts are concurrently used by multiple languages (e.g., overlapping usage between es and pt). This suggests that SRC-MoE separates language-specific acoustic–phonetic variation while retaining shared capacity for transferable structure. On the ST side (TGT-MoE, right), we observe an analogous but more target-oriented behavior: routing is target-dependent (e.g., de and en have different dominant experts), yet remains soft with noticeable weight on secondary experts, enabling partial sharing when beneficial. Such soft target-conditioned specialization provides a direct mechanism to reduce inter-direction interference in many-to-many translation without discarding sharing among related targets. In contrast, without entropy regularization, routing collapses into near-disjoint partitions: each expert becomes predominantly focused on a single language with little to no cross-language sharing, reducing the intended benefit of conditional capacity allocation across languages.

4 Conclusion

We propose LCMA-SRT, a unified many-to-many multilingual speech translation extension of hierarchical neural transducers. By inserting two lightweight language-conditional MoE adapters (SRC-MoE after the ASR encoder; TGT-MoE after the ST encoder), it allocates capacity across source and target languages while keeping a shared backbone. On Europarl-ST, LCMA-SRT markedly improves many-to-many translation over fully shared baselines and matches or surpasses direction-specific systems within a single model. Ablations and routing analysis show SRC-MoE mainly strengthens multilingual ASR representations, while TGT-MoE is crucial for stable target-controlled translation; together they provide complementary gains.

Limitations

This work evaluates LCMA-SRT only in the offline setting and does not benchmark streaming performance; moreover, experiments are primarily conducted on Europarl-ST with limited domain and language-family coverage. LCMA-SRT relies on explicit source/target language signals (and a target prefix for ST) to drive conditional routing, which may be affected by imperfect language identification or missing target specifications in practice. Finally, we do not systematically study accuracy–efficiency trade-offs or hyperparameter sensitivity (e.g., number of experts and entropy coefficient), leaving broader scalability and deployment considerations for future work.

References

Hilal Al Shamsi, Abdullah G Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: a systematic review. *Oman medical journal*, 35(2):e122.

Antonios Anastasopoulos and 1 others. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation*, pages 98–157.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *ICASSP 2018*, pages 6224–6228.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Nicola Bertoldi and Marcello Federico. 2005. A new decoder for spoken language translation based on confusion networks. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 86–91.

Chung-Cheng Chiu and 1 others. 2019. A comparison of end-to-end models for long-form speech recognition. In *ASRU 2019*, pages 889–896.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *NAACL 2021*, pages 1882–1896.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*.

Marco Gaido, Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Amir Hussein, Dan Povey, Cihan Xiao, Leibny Paola Garcia, Matthew Wiesner, and Sanjeev Khudanpur. 2025. Hent-srt: Hierarchical efficient neural transducer with self-distillation for joint speech recognition and translation. *arXiv preprint arXiv:2506.02157*.

Amir Hussein, Brian Yan, Antonios Anastasopoulos, Shinji Watanabe, and Sanjeev Khudanpur. 2023. Enhancing end-to-end conversational speech translation through target language context utilization. *arXiv preprint arXiv:2309.15686*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Mahaveer Jain, Kjell Schubert, Jay Mahadeokar, Ching-Feng Yeh, Kaustubh Kalgaonkar, Anuroop Sriram, Christian Fuegen, and Michael L. Seltzer. 2019. Rnn-t for latency controlled asr with improved beam search. *arXiv preprint arXiv:1911.01629*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Onur Köksal and Nurcihan Yürük. 2020. The role of translator in intercultural communication. *International Journal of Curriculum and Instruction*, 12(1):327–338.

Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey. 2022. Pruned rnn-t for fast, memory-efficient asr training. *arXiv preprint arXiv:2206.13236*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Junjie Li, Jing Peng, Yangui Fang, Shuai Wang, and Kai Yu. 2025. Mosa: Mixtures of simple adapters outperform monolithic approaches in llm-based multilingual asr. *arXiv preprint arXiv:2508.18998*.

Wei Liu, Jiahong Li, Yiwen Shao, and Dong Yu. 2025. Tta: Transcribe, translate and alignment for cross-lingual speech representation. *arXiv preprint arXiv:2511.14410*.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Interspeech*, pages 3177–3180.

B Mu, K Wei, Q Shao, Y Xu, and L Xie. 2024. Hdmole: Mixture of lora experts with hierarchical routing and dynamic thresholds for fine-tuning llm-based asr models. *arXiv preprint arXiv:2409.19878*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech*, pages 2613–2617.

Rastislav Rabatin, Frank Seide, and Ernie Chang. 2024. Navigating the minefield of mt beam search in cascaded streaming speech translation. *arXiv preprint arXiv:2407.11010*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Mark Sinclair, Peter Bell, Alexandra Birch, and Fergus McInnes. 2014. A semi-markov model for speech segmentation with an utterance-break prior. In *Interspeech*, pages 2351–2355.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96.

Yun Tang and 1 others. 2023. Hybrid transducer and attention based encoder-decoder modeling for speech-to-text tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12441–12455.

Peidong Wang, Eric Sun, Jian Xue, Yu Wu, Long Zhou, Yashesh Gaur, Shujie Liu, and Jinyu Li. 2023. Lamassu: A streaming language-agnostic multilingual speech recognition and translation model using neural transducers. In *Interspeech*, pages 57–61.

H Xue, W Ren, X Geng, K Wei, L Li, Q Shao, L Yang, K Diao, and L Xie. 2024. Ideal-llm: Integrating dual encoders and language-adapted llm for multilingual speech-to-text. *arXiv preprint arXiv:2409.11214*.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. Ctc alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

Zengwei Yao, Wei Kang, Xiaoyu Yang, Fangjun Kuang, Liyong Guo, Han Zhu, Zengrui Jin, Zhaoqing Li, Long Lin, and Daniel Povey. 2024. Cr-ctc: Consistency regularization on ctc for improved speech recognition. *arXiv preprint arXiv:2410.05101*.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2024. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *CoRR*, abs/2406.02166.

Piotr Żelasko, Daniel Povey, Jan Trmal, and Sanjeev Khudanpur. 2021. Lhotse: a speech data representation library for the modern deep learning ecosystem. *arXiv preprint arXiv:2110.12561*.

A Per-direction Results for Joint ASR and ST

SRC \ TGT	Model	WER (%)↓								
		de	**en**	**es**	**fr**	**it**	**nl**	**pl**	**pt**	**ro**
de	HENT-SRT-M2O×9	-	21.80	26.64	27.12	27.44	26.43	26.51	26.51	26.62
	HENT-SRT-M2M	-	19.09	18.77	18.82	19.09	18.86	18.79	18.99	18.86
	LCMA-SRT	-	**18.01**	**17.84**	**17.85**	**18.23**	**17.92**	**17.75**	**17.93**	**17.99**
en	HENT-SRT-M2O×9	16.42	-	17.32	17.56	17.08	17.45	17.29	17.18	17.21
	HENT-SRT-M2M	13.92	-	13.94	14.02	13.79	13.84	13.99	13.90	13.53
	LCMA-SRT	**12.94**	-	**12.93**	**13.02**	**12.84**	**12.87**	**12.92**	**12.97**	**12.64**
es	HENT-SRT-M2O×9	21.29	17.77	-	22.25	22.83	22.68	21.93	22.82	22.66
	HENT-SRT-M2M	15.96	15.80	-	15.91	15.69	15.97	15.85	15.89	15.75
	LCMA-SRT	**15.30**	**15.14**	-	**15.27**	**15.02**	**15.31**	**15.26**	**15.25**	**15.14**
fr	HENT-SRT-M2O×9	19.37	16.07	19.82	-	20.41	19.30	19.45	19.86	20.80
	HENT-SRT-M2M	13.40	13.38	13.28	-	13.42	13.36	13.39	13.38	13.37
	LCMA-SRT	**12.58**	**12.51**	**12.53**	-	**12.51**	**12.50**	**12.56**	**12.55**	**12.65**
it	HENT-SRT-M2O×9	18.18	15.05	19.19	19.32	-	19.06	18.60	19.00	19.91
	HENT-SRT-M2M	13.10	13.19	13.13	13.24	-	13.17	12.98	13.18	13.27
	LCMA-SRT	**12.50**	**12.41**	**12.52**	**12.63**	-	**12.59**	**12.42**	**12.62**	**12.66**
nl	HENT-SRT-M2O×9	38.99	32.95	38.85	38.85	39.52	-	38.99	39.32	39.26
	HENT-SRT-M2M	28.59	28.65	28.73	28.46	28.62	-	28.46	28.46	28.47
	LCMA-SRT	**27.01**	**27.23**	**26.89**	**26.91**	**27.20**	-	**26.93**	**27.07**	**26.82**
pl	HENT-SRT-M2O×9	25.89	22.01	26.33	27.19	25.99	26.47	-	27.13	27.36
	HENT-SRT-M2M	18.26	18.27	18.14	18.21	17.87	18.27	-	18.29	18.00
	LCMA-SRT	**17.54**	**17.39**	**17.32**	**17.36**	**17.01**	**17.43**	-	**17.57**	**17.11**
pt	HENT-SRT-M2O×9	19.90	16.27	21.74	20.82	20.77	20.99	20.48	-	20.53
	HENT-SRT-M2M	13.60	13.59	13.52	13.59	13.38	13.58	13.57	-	13.34
	LCMA-SRT	**12.37**	**12.72**	**12.28**	**12.37**	**12.08**	**12.38**	**12.40**	-	**12.19**
ro	HENT-SRT-M2O×9	22.32	15.85	21.87	22.04	23.97	22.88	23.63	22.82	-
	HENT-SRT-M2M	14.59	14.20	14.52	14.42	14.17	14.59	14.49	14.65	-
	LCMA-SRT	**13.64**	**13.29**	**13.51**	**13.46**	**13.38**	**13.61**	**13.54**	**13.72**	-

Table 4: Direction-wise ASR WER (%, ↓) on the Europarl-ST test set under the same (SRC,TGT)-conditioned decoding used in joint ASR and ST. Rows denote source speech languages and columns denote target translation languages. We compare HENT-SRT-M2O×9, HENT-SRT-M2M, and LCMA-SRT. Bold indicates the lowest WER for each direction.

SRC \ TGT	Model	BLEU ↑								
		de	**en**	**es**	**fr**	**it**	**nl**	**pl**	**pt**	**ro**
de	HENT-SRT-M2O×9	-	17.5	13.3	12.1	8.7	16.2	5.9	12.4	8.3
	HENT-SRT-M2M	-	11.0	3.7	3.3	1.1	4.1	1.6	4.0	2.2
	LCMA-SRT	-	**22.0**	**19.7**	**20.2**	**14.5**	**19.0**	**8.9**	**18.7**	**13.5**
en	HENT-SRT-M2O×9	15.4	-	26.0	24.6	19.0	21.9	9.7	23.1	19.8
	HENT-SRT-M2M	4.0	-	9.7	6.5	3.1	5.3	1.6	7.1	4.5
	LCMA-SRT	**20.1**	-	**33.4**	**30.7**	**25.0**	**25.4**	**14.7**	**29.4**	**26.3**
es	HENT-SRT-M2O×9	9.9	22.1	-	20.2	15.7	15.1	6.9	22.4	12.2
	HENT-SRT-M2M	2.1	13.4	-	3.9	1.5	3.1	0.9	5.4	2.2
	LCMA-SRT	**13.7**	**26.1**	-	**26.3**	**21.0**	**19.4**	**10.3**	**26.6**	**17.7**
fr	HENT-SRT-M2O×9	11.0	23.5	20.3	-	17.6	16.9	7.4	23.3	13.0
	HENT-SRT-M2M	2.9	11.9	6.4	-	2.2	4.0	1.3	6.5	2.4
	LCMA-SRT	**14.9**	**28.6**	**27.0**	-	**22.5**	**21.3**	**11.1**	**27.5**	**18.3**
it	HENT-SRT-M2O×9	11.3	23.0	21.3	20.3	-	16.1	8.3	22.4	13.4
	HENT-SRT-M2M	2.9	14.7	5.1	4.0	-	3.2	1.7	5.6	2.0
	LCMA-SRT	**14.8**	**27.0**	**27.3**	**25.3**	-	**20.2**	**11.0**	**26.1**	**17.8**
nl	HENT-SRT-M2O×9	7.1	15.6	11.3	10.4	7.3	-	3.7	10.4	6.3
	HENT-SRT-M2M	2.3	9.8	3.1	2.6	1.2	-	0.9	2.9	1.9
	LCMA-SRT	**12.1**	**21.0**	**17.6**	**16.5**	**13.6**	-	**7.0**	**16.9**	**11.6**
pl	HENT-SRT-M2O×9	9.5	19.3	17.1	15.7	11.9	14.3	-	14.6	10.0
	HENT-SRT-M2M	2.4	12.1	4.6	3.8	1.6	3.4	-	3.8	2.1
	LCMA-SRT	**14.3**	**23.9**	**24.1**	**22.9**	**18.6**	**19.5**	-	**20.8**	**16.5**
pt	HENT-SRT-M2O×9	10.9	23.7	22.1	21.3	17.3	15.6	7.5	-	13.9
	HENT-SRT-M2M	2.3	13.3	6.6	4.0	1.9	3.0	1.1	-	2.5
	LCMA-SRT	**15.4**	**28.1**	**28.3**	**27.0**	**22.8**	**19.7**	**10.5**	-	**19.0**
ro	HENT-SRT-M2O×9	10.9	25.3	21.4	21.4	15.8	16.0	7.9	18.8	-
	HENT-SRT-M2M	1.9	16.4	4.6	3.7	1.5	2.2	0.7	3.9	-
	LCMA-SRT	**15.8**	**30.1**	**28.9**	**28.4**	**22.1**	**19.7**	**12.2**	**25.3**	-

Table 5: Direction-wise speech translation BLEU scores on Europarl-ST test set (9 languages, 72 directions). Rows denote source speech languages and columns denote target translation languages. We compare the direction-specific hierarchical transducer baseline HENT-SRT-M2O×9, the fully shared many-to-many baseline HENT-SRT-M2M, and our unified LCMA-SRT. Bold indicates the best score for each direction.

SRC \ TGT	Model	COMET ↑								
		de	**en**	**es**	**fr**	**it**	**nl**	**pl**	**pt**	**ro**
de	HENT-SRT-M2O×9	-	0.615	0.531	0.479	0.504	0.549	0.521	0.544	0.545
	HENT-SRT-M2M	-	0.522	0.453	0.407	0.409	0.397	0.383	0.447	0.391
	LCMA-SRT	-	**0.683**	**0.624**	**0.572**	**0.591**	**0.604**	**0.591**	**0.636**	**0.627**
en	HENT-SRT-M2O×9	0.571	-	0.641	0.606	0.625	0.620	0.584	0.668	0.680
	HENT-SRT-M2M	0.421	-	0.533	0.470	0.487	0.430	0.419	0.524	0.458
	LCMA-SRT	**0.638**	-	**0.741**	**0.690**	**0.714**	**0.674**	**0.663**	**0.749**	**0.765**
es	HENT-SRT-M2O×9	0.488	0.652	-	0.548	0.571	0.534	0.546	0.636	0.589
	HENT-SRT-M2M	0.357	0.536	-	0.416	0.424	0.385	0.374	0.464	0.396
	LCMA-SRT	**0.544**	**0.708**	-	**0.627**	**0.657**	**0.584**	**0.609**	**0.709**	**0.663**
fr	HENT-SRT-M2O×9	0.499	0.685	0.603	-	0.603	0.551	0.555	0.650	0.618
	HENT-SRT-M2M	0.373	0.535	0.484	-	0.440	0.396	0.385	0.481	0.408
	LCMA-SRT	**0.561**	**0.737**	**0.700**	-	**0.685**	**0.603**	**0.616**	**0.723**	**0.701**
it	HENT-SRT-M2O×9	0.507	0.679	0.614	0.569	-	0.551	0.568	0.650	0.623
	HENT-SRT-M2M	0.372	0.560	0.477	0.425	-	0.393	0.380	0.472	0.404
	LCMA-SRT	**0.560**	**0.728**	**0.698**	**0.640**	-	**0.600**	**0.619**	**0.717**	**0.686**
nl	HENT-SRT-M2O×9	0.444	0.581	0.500	0.460	0.467	-	0.486	0.509	0.509
	HENT-SRT-M2M	0.367	0.508	0.435	0.397	0.402	-	0.365	0.435	0.380
	LCMA-SRT	**0.538**	**0.660**	**0.595**	**0.544**	**0.561**	-	**0.556**	**0.604**	**0.593**
pl	HENT-SRT-M2O×9	0.515	0.643	0.568	0.518	0.545	0.543	-	0.584	0.583
	HENT-SRT-M2M	0.385	0.539	0.469	0.424	0.429	0.397	-	0.462	0.401
	LCMA-SRT	**0.584**	**0.709**	**0.667**	**0.612**	**0.651**	**0.608**	-	**0.683**	**0.677**
pt	HENT-SRT-M2O×9	0.522	0.692	0.631	0.584	0.605	0.556	0.576	-	0.636
	HENT-SRT-M2M	0.381	0.557	0.491	0.439	0.444	0.402	0.390	-	0.409
	LCMA-SRT	**0.581**	**0.744**	**0.722**	**0.662**	**0.695**	**0.609**	**0.632**	-	**0.710**
ro	HENT-SRT-M2O×9	0.514	0.697	0.606	0.575	0.596	0.563	0.569	0.627	-
	HENT-SRT-M2M	0.381	0.585	0.487	0.443	0.446	0.408	0.386	0.480	-
	LCMA-SRT	**0.587**	**0.753**	**0.711**	**0.667**	**0.696**	**0.624**	**0.642**	**0.724**	-

Table 6: Direction-wise speech translation COMET scores on Europarl-ST test set (9 languages, 72 directions). Rows denote source speech languages and columns denote target translation languages. We compare the direction-specific hierarchical transducer baseline HENT-SRT-M2O×9, the fully shared many-to-many baseline HENT-SRT-M2M, and our unified LCMA-SRT. Bold indicates the best score for each direction.

SRC \ TGT	Model	LMR (%)↓								
		de	**en**	**es**	**fr**	**it**	**nl**	**pl**	**pt**	**ro**
de	HENT-SRT-M2O×9	-	**0.08**	0.70	0.64	0.66	1.00	**0.00**	3.39	**1.70**
	HENT-SRT-M2M	-	56.60	87.83	90.14	94.98	77.09	83.11	83.75	82.55
	LCMA-SRT	-	0.38	**0.49**	**0.43**	**0.58**	**0.84**	0.73	**2.38**	1.79
en	HENT-SRT-M2O×9	**0.00**	-	**0.79**	0.25	0.35	0.65	**0.16**	**1.98**	**1.64**
	HENT-SRT-M2M	78.21	-	78.93	86.08	95.93	78.54	85.14	81.62	80.55
	LCMA-SRT	0.08	-	0.95	**0.16**	**0.09**	**0.24**	0.40	**1.98**	**1.64**
es	HENT-SRT-M2O×9	**0.18**	**0.22**	-	**0.18**	**0.46**	1.01	**0.09**	**1.19**	**0.88**
	HENT-SRT-M2M	88.51	58.54	-	92.98	96.76	89.58	90.93	80.07	86.81
	LCMA-SRT	0.54	0.61	-	0.37	0.65	**0.92**	0.38	1.29	1.54
fr	HENT-SRT-M2O×9	**0.00**	**0.11**	**0.55**	-	**0.00**	**0.26**	**0.18**	**1.55**	**1.16**
	HENT-SRT-M2M	85.18	71.51	87.89	-	95.60	86.66	89.85	77.71	84.93
	LCMA-SRT	**0.00**	0.33	0.73	-	0.38	0.87	0.36	1.91	**1.16**
it	HENT-SRT-M2O×9	**0.11**	**0.00**	**0.57**	**0.23**	-	**0.36**	**0.25**	2.20	**0.54**
	HENT-SRT-M2M	88.54	57.23	92.05	94.25	-	91.61	92.77	86.59	90.38
	LCMA-SRT	**0.11**	0.42	0.91	**0.23**	-	0.84	0.37	**1.62**	0.95
nl	HENT-SRT-M2O×9	**0.09**	**0.34**	**0.49**	**0.49**	0.90	-	**0.21**	3.18	**1.60**
	HENT-SRT-M2M	78.25	56.62	88.85	89.02	93.71	-	86.96	82.91	81.98
	LCMA-SRT	0.19	0.86	1.19	0.79	**0.79**	-	0.62	**1.80**	1.94
pl	HENT-SRT-M2O×9	**0.00**	**0.22**	**0.72**	**0.16**	**0.42**	1.80	-	2.08	1.72
	HENT-SRT-M2M	82.54	60.55	89.70	91.34	96.01	86.68	-	83.45	87.29
	LCMA-SRT	**0.00**	0.40	0.80	0.48	0.85	**1.14**	-	**1.60**	**1.01**
pt	HENT-SRT-M2O×9	**0.00**	**0.17**	0.16	**0.16**	**0.08**	**0.41**	**0.08**	-	**0.81**
	HENT-SRT-M2M	88.36	65.79	86.39	91.99	97.51	90.31	90.80	-	87.09
	LCMA-SRT	0.16	0.39	**0.08**	0.24	**0.08**	0.81	0.50	-	1.17
ro	HENT-SRT-M2O×9	**0.00**	**0.15**	**0.42**	0.26	**0.77**	**0.91**	**0.00**	1.58	-
	HENT-SRT-M2M	90.98	50.74	93.27	95.33	98.46	93.14	93.04	89.42	-
	LCMA-SRT	0.16	0.56	0.50	**0.17**	**0.77**	1.24	0.43	**1.33**	-

Table 7: Direction-wise ST LMR (%, ↓) on the Europarl-ST test set (9 languages, 72 directions). Rows denote source speech languages and columns denote target translation languages. We compare HENT-SRT-M2O×9, HENT-SRT-M2M, and LCMA-SRT. Bold indicates the lowest LMR for each direction.