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

Nanjie Li[1,2]*, **Xiaoyong Guo**[1,2]*
Hao Huang[1,3]†, **Xu Haihua**[2], **Wei Shi**[2]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Timekettle AI Lab, Shenzhen, China
[3]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China

Abstract

Neural transducers offer an alignment-free framework for speech-to-text modeling, and hierarchical transducer architectures further improve multilingual joint automatic speech recognition (ASR) and speech translation (ST) by stacking a translation-focused encoder on top of an ASR encoder. However, extending 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 for each 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) uses source-language embeddings to reduce cross-language interference and improve multilingual ASR. A target-conditioned MoE adapter (TGT-MoE) uses the desired target language to reduce cross-target interference and stabilize target-language generation 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 improving BLEU and COMET over strong hierarchical transducer baselines. We release our code and models at https://github.com/linanjie0820/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, in which an ASR module first transcribes speech into text and an MT system then translates the transcription (Matusov et al., 2005; Bertoldi and Federico, 2005). Although this modular design is simple and effective, cascaded systems can 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 ST has therefore emerged as a streamlined alternative that directly maps speech to target text (Bérard et al., 2016; Berard et al., 2018).

Most end-to-end ST work has been built on attention-based encoder–decoder (AED) models (Gaido et al., 2020). Beyond bilingual direct ST, prior work has extended this paradigm to multilingual and joint speech-to-text settings, including multilingual end-to-end ST (Inaguma et al., 2019; Di Gangi et al., 2019) and joint ASR+multilingual ST architectures such as the dual-decoder Transformer (Le et al., 2020). These studies show that cross-lingual and cross-task transfer can be beneficial, but they are mainly based on AED formulations.

Compared with AED models, alignment-free objectives such as CTC and neural transducers provide a frame-synchronous alternative that has been highly effective for speech modeling (Graves et al., 2006; Graves, 2012). Recent work has also extended neural transducers to end-to-end speech translation, including large-scale streaming and multilingual settings (Xue et al., 2022). However, ST is typically less monotonic than ASR and often requires substantial reordering (Yan et al., 2023). As a result, jointly modeling ASR and ST with fully shared transducer representations can be challenging, especially for language pairs with strong word-order divergence (Tang et al., 2023; Hussein et al., 2025). Hierarchical transducer architectures mitigate this mismatch by explicitly separating recognition-oriented and translation-oriented

*This work was completed during the internship.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35363–35377

processing within one model. In particular, HENT-SRT (Hussein et al., 2025) first produces speech-aligned, approximately monotonic representations with an ASR encoder and then applies a translation-focused encoder to better model non-monotonic reordering. While this design improves joint ASR and ST, scaling hierarchical transducers to multilingual many-to-many settings remains difficult: fully shared multilingual models can suffer from negative transfer, language imbalance, and unstable target-language generation, whereas training separate models for each direction is computationally impractical. Large multilingual speech systems such as Whisper, Seamless, and Canary also benefit from scale (Radford et al., 2023; Barrault et al., 2023; Sekoyan et al., 2025), but are typically studied under substantially larger model and compute budgets than the transducer setting considered here.

A related line of work studies parameter-efficient specialization for multilingual speech-to-text models. Efficient finetuning of pretrained models has been shown effective for multilingual ST (Li et al., 2021), and lightweight adapters have been used to specialize multilingual ST models to particular language pairs with modest parameter overhead (Le et al., 2021). For streaming multilingual ASR, Bai et al. (Bai et al., 2024) further show that lightweight language-dependent adapters can substantially improve tail-language performance under multilingual imbalance. These findings suggest that lightweight conditional modules are a promising way to retain shared multilingual representations while introducing language-specific specialization.

Motivated by these observations, we propose LCMA-SRT, a unified multilingual hierarchical neural transducer that augments a hierarchical transducer backbone (Hussein et al., 2025) with language-conditional Mixture-of-Experts (MoE) adapters for many-to-many joint ASR and ST. LCMA-SRT is designed to address the central challenge of multilingual many-to-many transduction, where 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. Specifically, we introduce two complementary language-conditional adapters: a source-conditioned MoE adapter (SRC-MoE), which improves multilingual ASR by reducing cross-language interference in speech-aligned representations, and a target-conditioned MoE adapter (TGT-MoE), which reduces cross-target interference and stabilizes target-language generation for many-to-many ST. This design extends hierarchical transducer modeling from many-to-one settings to a single unified many-to-many model for joint multilingual ASR and ST, while enabling source-side and target-side specialization without replicating the shared backbone. Experiments on Europarl-ST demonstrate that LCMA-SRT improves both ASR and ST over matched hierarchical transducer baselines within one joint model, and additionally outperforms Whisper Base on the Europarl-ST X→En subset. Controlled ablations further confirm the contribution of each design component in many-to-many multilingual transduction.

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 $\mathbf{a} \in \bar{\mathcal{V}}_{\mathrm{T+U_k}}^{(k)}$:

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

where $\bar{\mathcal{V}}^{(k)} = \mathcal{V}^{(k)} \cup \{\phi\}$ augments the task vocabulary with the blank symbol ϕ, 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:

$$\mathbf{F}^{(s)} = \mathrm{Enc}_{\mathrm{asr}}(X) \quad (2)$$

$$\mathbf{F}^{(t)} = \mathrm{Enc}_{\mathrm{st}}\left(\mathbf{F}^{(s)}\right) \quad (3)$$

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

$$\mathcal{L}_{\mathrm{nt}} = \alpha_{\mathrm{asr}} \mathcal{L}_{\mathrm{nt}}^{(s)} + \alpha_{\mathrm{st}} \mathcal{L}_{\mathrm{nt}}^{(t)} \quad (4)$$

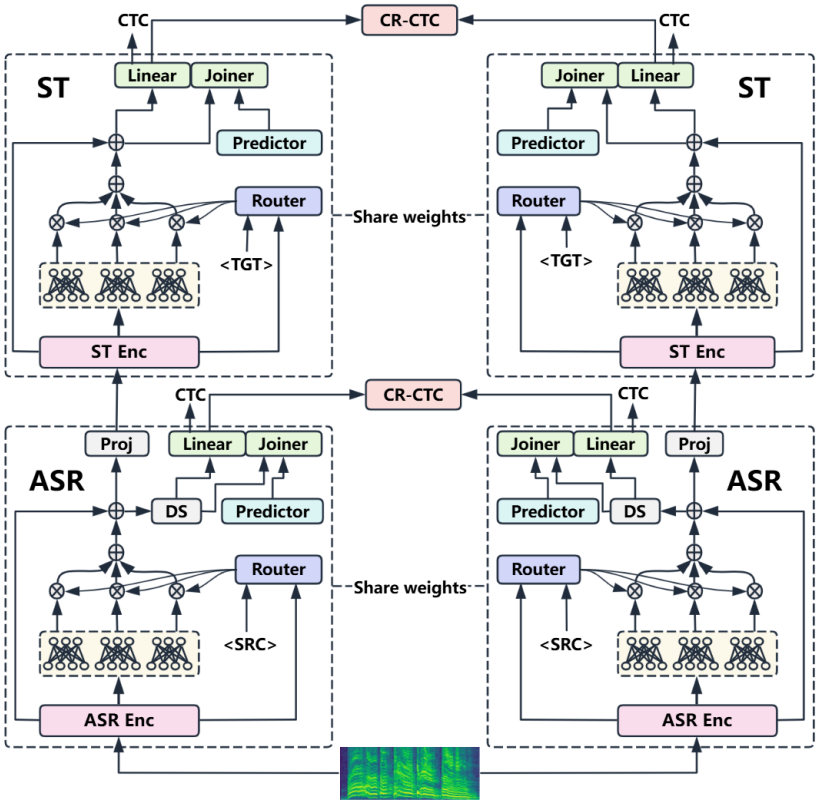

Figure 1: Proposed language-conditional mixture-of-experts adapters for unified multilingual speech recognition and translation (LCMA-SRT). SRC-MoE is inserted after the ASR encoder to specialize speech-aligned multilingual representations, while TGT-MoE is inserted after the ST encoder to specialize translation-oriented representations. Both tasks use task-specific predictor–joiner transducer heads, and ST decoding is further conditioned on a target-language prefix token.

where $\mathcal{L}_{\text{nt}}^{(k)} = -\log P(\mathbf{y}^{(k)} \mid X)$ and α_{asr}, α_{st} control the contribution of each task. We next describe how LCMA-SRT augments this hierarchical transducer with source-conditioned and target-conditioned MoE adapters for many-to-many multilingual ASR and ST.

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. 2) 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 $\mathbf{e}_s = \text{Embed}_{\text{src}}(\ell_s) \in \mathbb{R}^{d_\ell}$ 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).

We place SRC-MoE at the output of the ASR encoder because this boundary concentrates multilingual speech variation while keeping the added computation localized to a single lightweight module. In the hierarchical transducer, it therefore provides a practical efficiency–stability trade-off: earlier insertion would affect more layers and increase cost, whereas later insertion would miss the most direct point for modulating the shared speech-aligned representation before it is consumed by both the ASR head and the downstream ST branch.

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

$$\mathbf{w}_t = \text{softmax}(g([\mathbf{h}_t; \mathbf{e}_s])) \in \mathbb{R}^E \tag{5}$$

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:

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t + \sum_{i=1}^{E} w_{t,i}\, f_i(\mathbf{h}_t) \tag{6}$$

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 $2\times$ downsampling module before the ASR transducer head, yielding downsampled representations $\hat{\mathbf{F}}^{(s)}$. 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 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 $\tilde{\mathbf{F}}^{(s)}$ before feeding it into the ST encoder, which then produces translation-oriented representations:

$$\mathbf{F}^{(t)} = \mathrm{Enc}_{\mathrm{st}}\Big(\mathrm{Proj}\Big(\tilde{\mathbf{F}}^{(s)}\Big)\Big) \tag{7}$$

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. 5–6, where the hidden input is set to $\mathbf{h}_t = \mathbf{f}_t^{(t)}$ and the conditioning embedding is $\mathbf{e} = \mathbf{e}_t$. We place TGT-MoE at the output of the ST encoder because this is the most direct boundary at which language-conditioned specialization can act on translation-oriented representations rather than raw acoustic features. In our setting, this design directly targets target-language drift in the many-to-many setting while keeping the shared hierarchical backbone unchanged. The adapted ST representations are

$$\tilde{\mathbf{F}}^{(t)} = \mathrm{Adapter}_{\mathrm{tgt}}\Big(\mathbf{F}^{(t)}; \mathbf{e}_t\Big) \tag{8}$$

Conditioning routing on the target language reduces cross-target interference and stabilizes target-language generation in many-to-many ST. We do not claim that this placement is globally optimal; rather, it is a practical boundary choice that balances interpretability, efficiency, and stability in the hierarchical transducer. 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 $\hat{\mathbf{F}}^{(s)}$; for ST, they are computed on the TGT-MoE output $\tilde{\mathbf{F}}^{(t)}$ (Eq. 8). For each task $k \in \{s, t\}$, we define the auxiliary loss as

$$\mathcal{L}_{\mathrm{aux}}^{(k)} = \lambda_{\mathrm{ctc}}^{(k)} \mathcal{L}_{\mathrm{ctc}}^{(k)} + \lambda_{\mathrm{cr}}^{(k)} \mathcal{L}_{\mathrm{cr}}^{(k)} \tag{9}$$

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 $\lambda_{\mathrm{ctc}}^{(k)}$ and $\lambda_{\mathrm{cr}}^{(k)}$ weight the CTC and CR-CTC terms for each task.

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 $\bar{H}(\mathbf{w})$ denote the mean per-frame entropy of the routing distribution over valid (non-padded) frames. We apply entropy regularization to both SRC-MoE and TGT-MoE, and average the two weighted terms:

$$\mathcal{L}_{\mathrm{ent}} = -\frac{1}{2} \sum_{k \in \{\mathrm{s,t}\}} \lambda_{\mathrm{ent}}^{(k)} \cdot \bar{H}\Big(\mathbf{w}^{(k)}\Big) \tag{10}$$

Here $\lambda_{\mathrm{ent}}^{(\mathrm{s})}$ and $\lambda_{\mathrm{ent}}^{(\mathrm{t})}$ 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:

$$\mathcal{L} = \mathcal{L}_{\mathrm{nt}} + \alpha_{\mathrm{asr}} \mathcal{L}_{\mathrm{aux}}^{(s)} + \alpha_{\mathrm{st}} \mathcal{L}_{\mathrm{aux}}^{(t)} + \mathcal{L}_{\mathrm{ent}} \tag{11}$$

In all experiments, we keep the multitask and auxiliary weights fixed across runs rather than tuning them separately for each model. We set both α_{asr} and α_{st} to 1, use 0.1 for $\lambda_{\text{ctc}}^{(s)}$ and $\lambda_{\text{ctc}}^{(t)}$, and 0.05 for $\lambda_{\text{cr}}^{(s)}$ and $\lambda_{\text{cr}}^{(t)}$. When MoE adapters are enabled, we set both $\lambda_{\text{ent}}^{(\text{s})}$ and $\lambda_{\text{ent}}^{(\text{t})}$ to 0.015.

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). In our implementation, both tasks use a stateless predictor implemented as a single Conv1D layer with kernel size 2, together with a standard joiner. During training, the ASR and ST predictor–joiner heads are optimized jointly with the objectives above; during inference, beam search is applied on the corresponding task head to produce the final output sequence. 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 by itself prevent target-language drift in fully shared many-to-many models (Sec. 3.2), which further motivates the use of target-conditioned routing.

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 use Icefall[1] and Lhotse (Żelasko et al., 2021). ASR pretraining uses 276 hours of unique audio, and joint ASR/ST training reuses the same segments with all available target translations, yielding 1,642 pair-hours. We resample all audio to 16 kHz and extract 80-dimensional log-Mel filterbank features with a 25 ms window and 10 ms frame shift. We apply on-the-fly SpecAugment (Park et al., 2019); for CR-CTC self-distillation, we strengthen the second view following (Yao et al., 2024). For text, we apply Whisper-style normalization (Radford et al., 2023) and train 1k SentencePiece BPE models (Kudo and Richardson, 2018) for both ASR and ST. ASR uses a shared multilingual BPE model; ST uses a shared multilingual 1k BPE model with a prepended target-language tag for many-to-many systems, while HENT-SRT-M2O×9 uses one 1k ST BPE model per target language.

Models: We use the same hierarchical Zipformer transducer backbone as HENT-SRT for fair comparison, with an ASR encoder stacked below an ST encoder and task-specific transducer heads. 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, while adding SRC-MoE increases the parameter count to 36M.

For Stage 2 joint ASR and ST training, we compare (i) HENT-SRT-M2O×9, i.e., nine separate many-to-one models (61M each; 549M in total), (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 (77M) together with its ablations. The size-matched unconditioned control TGT-MoE→MoE also has 77M parameters, while TGT-MoE→T-Bias and w/o TGT-MoE each have 66M parameters, and w/o SRC-MoE has 71M parameters. Unless otherwise specified, we use E=8 experts for SRC-MoE and E=16 experts for TGT-MoE. Both tasks use a stateless transducer predictor implemented as a single Conv1D layer with kernel size 2, together with a standard joiner.

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

To strengthen baseline positioning beyond hierarchical transducer comparisons, we additionally report Whisper Base on the Europarl-ST X→En subset (8 source languages), which serves as a compact external multilingual baseline under a similar parameter scale (74M vs. 77M for LCMA-SRT). We treat this comparison as complementary to the matched hierarchical-transducer baselines above, since Whisper is trained under a different objective and recipe.

We optimize the transducer objective with the pruned transducer loss (Kuang et al., 2022) and use 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.

Evaluation: To ensure a comprehensive evaluation of speech recognition and translation quality, we utilize BLEU (Papineni et al., 2002) for surface-level matching, COMET (Rei et al., 2020) for semantic adequacy, and sentence-level target-language mismatch rate (LMR) using an off-the-shelf language identification model (Joulin et al., 2016), where a hypothesis is counted as matched only if it is classified as the specified target language with confidence ≥ 0.7. Since ST targets are normalized during training and decoding, translation is evaluated in the same normalized form. ASR performance is assessed using word error rate (WER).

3.2 Main Results

We evaluate LCMA-SRT on Europarl-ST following the two-stage training recipe. We first analyze multilingual ASR pretraining and then compare 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. This suggests that source-conditioned routing learns more robust speech-aligned representations for subsequent joint ASR/ST training.

Many-to-Many Joint Training: We present the many-to-many joint training results in Tables 2 and 3. Compared with the direction-specific baseline HENT-SRT-M2O×9, LCMA-SRT improves average BLEU by +5.2 (15.3 → 20.5) and COMET by +0.076 (0.575 → 0.651), while reducing average WER from 23.28% to 15.71%. Notably, these gains are achieved with a single 77M model, compared with nine separate HENT-SRT-M2O systems (549M in total), indicating that LCMA-SRT improves both translation quality and recognition accuracy with a substantially smaller overall model footprint. We also observe consistent gains across target languages, suggesting effective transfer across directions within one multilingual model.

Comparison with the fully shared baseline HENT-SRT-M2M further highlights the importance of explicit target conditioning. Despite forced target-prefix decoding, HENT-SRT-M2M suffers from severe cross-direction interference and target-language drift (84.95% LMR), resulting in poor translation quality (4.3 BLEU / 0.436 COMET). In contrast, LCMA-SRT reduces LMR to 0.75%, close to HENT-SRT-M2O×9 (0.65%), while simultaneously improving BLEU, COMET, and WER. These results show that hierarchical transducers can be effectively extended to the many-to-many setting when source-side and target-side conditioning are introduced explicitly. Detailed direction-wise results are reported in Appendix A: ASR WER in Table 5, and ST BLEU, COMET, and LMR in Tables 6, 7, and 8.

Comparison to Whisper Base: We additionally compare LCMA-SRT with Whisper Base on the Europarl-ST X→En subset to strengthen baseline positioning. At a similar parameter scale (77M vs. 74M), LCMA-SRT substantially outperforms Whisper Base on AST, achieving 25.9 BLEU / 0.715 COMET versus 15.6 BLEU / 0.626 COMET, and also improves ASR performance (16.09 vs. 23.66 WER on average). Detailed results are reported in Appendix A, Table 4.

3.3 Ablation Analysis

We analyze the ablations in Tables 2 and 3 to assess the contribution of each component in LCMA-SRT. We first examine TGT-MoE. Removing TGT-MoE (w/o TGT-MoE) severely degrades translation quality, reducing BLEU to 4.2 and causing extremely high target-language mismatch (85.19% LMR), showing that a monolithic shared ST en-

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	0.75	**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.

coder cannot preserve target-language fidelity in the 72-direction many-to-many setting. Replacing TGT-MoE with a size-matched unconditioned MoE (TGT-MoE→MoE, 77M) yields similarly high LMR (85.23%) together with poor BLEU and COMET, indicating that the gains do not come from parameter count alone, but from language-conditional routing. A target-identity bias variant (TGT-MoE→T-Bias) restores low LMR (0.78%), showing that explicit directional cues already help prevent language leakage in many-to-many translation. However, it still underperforms the full TGT-MoE model in both BLEU and COMET, suggesting that target-language cues are necessary but not sufficient: beyond selecting the correct output language, conditional expert capacity also improves translation quality within that language.

On the ASR side, removing SRC-MoE (w/o SRC-MoE) increases average WER from 15.71% to 16.11% and reduces ST BLEU from 20.5 to 20.0. This indicates that SRC-MoE mainly strengthens multilingual ASR representations while also providing a modest downstream benefit for ST. Overall, the two adapters play complementary roles: SRC-MoE mainly improves source-side representations, whereas TGT-MoE is the main driver of target-language fidelity and many-to-many ST stability.

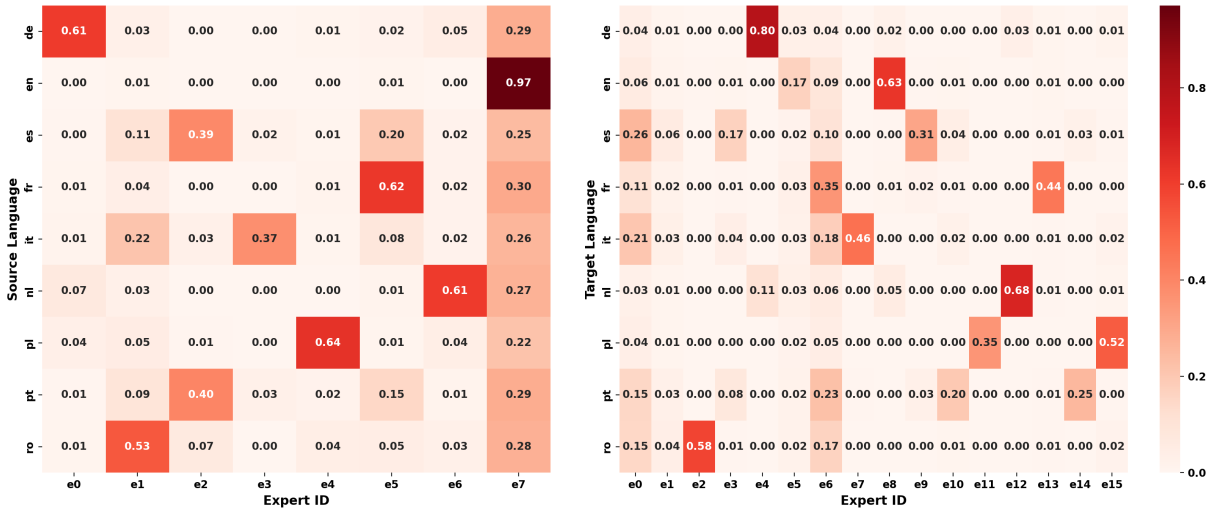

Figure 2: Heatmaps of language-aggregated expert routing weights in the language-conditional MoE adapters. Left: SRC-MoE on the ASR side, where each row is a source language and each column is an expert (e0–e7). Right: TGT-MoE on the ST side, where each row is a target language and each column is an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

3.4 Sensitivity Analysis

We conduct small sensitivity studies on expert counts and the entropy regularization coefficient. Within the tested range, the default setting $(E_{\text{SRC}}, E_{\text{TGT}}) = (8, 16)$ with $\lambda_{\text{ent}} = 0.015$ provides the best overall balance across LMR, WER, BLEU, and COMET. Increasing E_{TGT} from 8 to 16 yields the main improvement, while further increasing it to 20 shows diminishing returns. In contrast, reducing E_{SRC} from 8 to 4 has a smaller effect on translation quality but slightly harms WER. Performance is also reasonably stable across $\lambda_{\text{ent}} \in \{0, 0.01, 0.015, 0.02\}$, with the best overall trade-off observed at 0.015. Full results are reported in Appendix A, Tables 9 and 10.

3.5 Analysis of MoE Routing Behavior

Figure 2 shows the language-aggregated routing distributions learned by the 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 shared across 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 pattern: routing is target-dependent (e.g., en and ro 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 plausible explanation for the observed reduction in inter-direction interference in many-to-many translation without discarding sharing among related targets. By contrast, without entropy regularization, routing becomes much more polarized, with each expert focusing predominantly on a single language and exhibiting little cross-language sharing. This qualitative behavior is consistent with the sensitivity results reported in Appendix A, Tables 9 and 10, where moderate entropy regularization provides the best quality–stability trade-off.

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 and target language signals to enable language-conditional routing. While this design is effective in the current setting, its robustness under uncertain or unavailable language specifications remains to be studied. Finally, while we include limited sensitivity studies on the number of experts and the entropy coefficient, we do not yet provide broader analyses of accuracy–efficiency trade-offs, insertion-position sensitivity, or deployment-oriented latency characteristics.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62466055) and the Research Project of the State Language Commission of China (Grant No. ZDI145-96).

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.

Junwen Bai, Bo Li, Qiujia Li, Tara N Sainath, and Trevor Strohman. 2024. Efficient adapter finetuning for tail languages in streaming multilingual asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10841–10845. IEEE.

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 Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

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.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 585–592. IEEE.

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.

Amir Hussein, Cihan Xiao, Matthew Wiesner, Dan Povey, Leibny Paola Garcia Perera, and Sanjeev Khudanpur. 2025. Hent-srt: Hierarchical efficient neural transducer with self-distillation for joint speech recognition and translation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 153–164.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

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.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838.

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

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*.

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, Anna Sun, Hirofumi Inaguma, Xinyue Chen, Ning Dong, Xutai Ma, Paden Tomasello, and Juan Pino. 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 (Volume 1: Long Papers)*, pages 12441–12455.

Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur. 2022. Large-scale streaming end-to-end speech translation with neural transducers. *arXiv preprint arXiv:2204.05352*.

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*.

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 Additional Results

Model	Param	WER↓	BLEU↑	COMET↑
HENT-SRT-M2O×9	549M	19.72	21.2	0.656
HENT-SRT-M2M	61M	17.02	12.8	0.543
Whisper Base	74M	23.66	15.6	0.626
LCMA-SRT	77M	**16.09**	**25.9**	**0.715**

Table 4: Comparison on the Europarl-ST X→En subset (8 source languages). We report average ASR WER, AST BLEU, and AST COMET. LCMA-SRT substantially outperforms Whisper Base on translation quality and also improves over Whisper Base and HENT-SRT-M2M in ASR.

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 5: 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 6: 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 7: 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 8: 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.

$(E_{\mathrm{SRC}}, E_{\mathrm{TGT}})$	**LMR↓**	**WER↓**	**BLEU↑**	**COMET↑**
$(8, 8)$	0.79	15.80	19.68	0.637
$(8, 12)$	0.90	**15.68**	20.06	0.643
$(8, 16)$	0.75	15.71	20.47	**0.651**
$(8, 20)$	0.87	15.69	**20.50**	0.647
$(4, 16)$	**0.73**	15.99	19.97	0.643

Table 9: Sensitivity to expert counts with $\lambda_{\mathrm{ent}} = 0.015$. The default setting $(E_{\mathrm{SRC}}, E_{\mathrm{TGT}}) = (8, 16)$ provides the best overall balance across LMR, WER, BLEU, and COMET in the tested range.

λ_{ent}	**LMR↓**	**WER↓**	**BLEU↑**	**COMET↑**
0	**0.74**	**15.64**	20.10	0.645
0.01	0.89	15.96	20.43	0.647
0.015	0.75	15.71	**20.47**	**0.651**
0.02	0.90	15.67	20.26	0.645

Table 10: Sensitivity to entropy regularization with $(E_{\mathrm{SRC}}, E_{\mathrm{TGT}}) = (8, 16)$. Performance is reasonably stable across the tested range, with $\lambda_{\mathrm{ent}} = 0.015$ giving the best overall balance.