

EDWARD-NMT: Elastic Dimensions With Adaptive Routing and Dual-consistency for Neural Machine Translation

Edward Nguyen*, Thai Tran*, Kiet Anh Hoang*, Duy Hoang*, Duy Vo, Khanh Le, Vu Le

UNEY, Switzerland

{edward.nguyen, thai.tt, kiet.ha, duy.hv, duy.vhb, khanh.ld, vu.lt}@uney.com

Abstract

Sparse Mixture-of-Experts (MoE) is a powerful paradigm for scaling Neural Machine Translation (NMT) and managing language heterogeneity. However, practical limitations hinder its deployment on resource-constrained devices: the entire expert pool must remain in memory, and the per-token compute budget is statically fixed during training by rigid top- k routing. We introduce **EDWARD-NMT** (Elastic Dimensions With Adaptive Routing and Dual-consistency for NMT), a novel MoE encoder-decoder architecture designed for a “train once, deploy anywhere” paradigm that exposes a family of deployable sub-networks at inference time. EDWARD-NMT integrates three key innovations: (i) Matryoshka-style elastic dimensions to flexibly reduce the memory footprint; (ii) an adaptive sigmoid router with a learned, task-conditioned threshold, allowing each token to dynamically determine its active expert count; and (iii) a dual-consistency objective to maximize the performance of the extracted sub-networks. Extensive experiments on PhoMT (En $\leftrightarrow$ Vi), WMT-14 (En $\leftrightarrow$ De), multilingual OPUS-16, and a multi-domain De $\leftrightarrow$ En benchmark validate our approach. The full-capacity architecture matches or exceeds strong MoE baselines with improved parameter efficiency. Furthermore, the sub-networks seamlessly extracted from a single checkpoint consistently outperform comparably sized dense Transformers without any re-training.

1 Introduction

Managing language and domain heterogeneity is a fundamental challenge in Neural Machine Translation (NMT) (Britz et al., 2017; Koehn and Knowles, 2017; Pham et al., 2021), as training corpora routinely mix diverse lexical and stylistic distributions (Stahlberg, 2020; Gain et al., 2026). The Mixture-

of-Experts (MoE) paradigm addresses this by routing tokens to conditionally activated experts, scaling capacity to specialize in distinct patterns while keeping per-token computation tractable (Shazeer et al., 2017; Fedus et al., 2022; Dai et al., 2022; Zhang et al., 2024; Liang et al., 2025).

Despite state-of-the-art performance, MoEs face deployment bottlenecks in resource-constrained environments. First, the entire expert pool must reside in memory (Hoffmann et al., 2022; Xu et al., 2023). While lightweight adapters reduce this overhead (Zhang et al., 2024), they still yield static, hardware-specific models. Second, standard MoEs rely on rigid top- k routing, applying an identical compute budget regardless of token complexity (Jawahar et al., 2023; Guo et al., 2025). Although Neural Architecture Search can design heterogeneous sub-networks (Jawahar et al., 2023), it requires complex offline phases. Consequently, NMT practitioners lack a unified framework to dynamically adjust memory and compute footprints on the fly.

To address these gaps, we propose **EDWARD-NMT** (Elastic Dimensions With Adaptive Routing and Dual-consistency for NMT), an encoder-decoder architecture enabling a *train-once, deploy-anywhere* paradigm through two coordinated forms of adaptivity. To overcome memory bottlenecks, we apply Matryoshka representation learning (Kusupati et al., 2022) to expert dimensions, allowing the MoE to be deterministically sliced into weight-sharing sub-networks at inference. To eliminate rigid compute constraints, we replace top- k selection with an adaptive sigmoid router using a learned, task-conditioned threshold, empowering tokens to dynamically dictate their active expert count. Finally, to ensure aggressively truncated sub-networks maintain translation quality without post-hoc fine-tuning, we introduce a novel in-place *dual-consistency* objective that explicitly aligns predictive distributions across nested structural scales.

Extensive experiments on PhoMT

*Equal contribution.

(En $\leftrightarrow$ Vi) (Doan et al., 2021), WMT-14 (En $\leftrightarrow$ De) (Bojar et al., 2014), multilingual OPUS-16 (Zhao et al., 2024), and a multi-domain De $\leftrightarrow$ En (Aharoni and Goldberg, 2020) benchmark validate our approach. EDWARD-NMT achieves strong full-capacity performance, improving upon Transformer Big (Vaswani et al., 2017) across benchmarks, while keeping smaller variants highly capable. Our full model matches recent lightweight MoEs (Zhang et al., 2024), while our smallest variant (EDWARD-NMT-S) outperforms a dense Transformer Base, with all variants seamlessly extracted from a single checkpoint. In summary, our main contributions are:

- We propose **EDWARD-NMT**, coupling elastic expert dimensions with dynamic, token-level sigmoid routing to realize a *train-once, deploy-anywhere* NMT architecture.
- We formulate a *dual-consistency* self-distillation objective, an entirely in-place regularization mechanism that stabilizes aggressively sliced sub-networks without requiring post-hoc adaptation or fine-tuning.
- We demonstrate that a single training run yields a discrete family of high-performing sub-networks, with the full-capacity model matching strong MoE baselines and smaller variants outperforming comparably sized dense networks.

2 Related Work

Static vs. Elastic Mixture-of-Experts. Sparse MoEs dominate sequence model scaling (Shazeer et al., 2017; Fedus et al., 2022; Dai et al., 2022; Jiang et al., 2024; Liu et al., 2024) and multi-domain NMT (Costa-Jussà et al., 2022; Zhang et al., 2024), yet produce static models with fixed memory footprints. Post-hoc compression (Lagunas et al., 2021; Li et al., 2025) yields rigid, hardware-specific models, while multi-sub-network architectures (Cai et al., 2020; Yu et al., 2019; Hou et al., 2020) require explicit fine-tuning. Existing elastic designs either restrict flexibility to dense networks (Kudugunta et al., 2024) or limit it to token-level granularity (Cappellazzo et al., 2026). While concurrent work (MoSE (Tashtan et al., 2026)) introduces slimmable experts, it retains rigid top- k routing and employs test-time training for width adaptation. In contrast,

EDWARD-NMT embeds Matryoshka-style elasticity directly into sparse experts. By introducing a variance-preservation scaling factor, we stabilize activations across nested widths, allowing a single trained MoE to yield a family of deployable sub-networks without any post-hoc adaptation.

Dynamic Routing and Adaptive Compute. Traditional MoEs employ rigid top- k routing, applying identical compute budgets regardless of token complexity and relying on restrictive load-balancing losses (Clark et al., 2022). Relaxations include soft (Puigcerver et al., 2024) and expert-choice routing (Zhou et al., 2022), while recent methods adapt per-token compute dynamically via null experts (AdaMoE (Zeng et al., 2024)), auto-tuned activations (DynMoE (Guo et al., 2025)), or differentiable ReLU routing (ReMoE (Wang et al., 2025)). EDWARD-NMT is concurrent but distinct: by coupling elastic experts with an adaptive sigmoid router, our framework uniquely enables dynamic adjustment of both memory and compute footprints at inference.

Self-distillation and Consistency Regularization. To enhance model generalization and robustness, typical consistency regularization aligns predictions across perturbed forward passes or augmented data views (Xie et al., 2020; Liang et al., 2021; Yao et al., 2025), while self-distillation transfers knowledge between layers or temporal states (Furlanello et al., 2018; Lan et al., 2018; Mobahi et al., 2020). Although in-place distillation successfully trains dense elastic networks by aligning sub-networks with a full-capacity teacher (Yu and Huang, 2019; Cai et al., 2020), it remains confined to statically routed architectures. We extend this via a novel *dual-consistency objective* tailored for elastic MoEs. Rather than aligning augmented views, we dynamically transfer knowledge between nested widths using a symmetric Kullback-Leibler (Jefreys) divergence. This bidirectional regularization anchors the full-capacity network while preserving the quality of aggressively truncated variants.

Shared Experts. To prevent sparse experts from redundantly learning general linguistic patterns, architectures like DeepSeek-MoE (Dai et al., 2024) and Qwen3 (Yang et al., 2025) isolate common knowledge into a dense, universally shared expert. EDWARD-NMT extends this by dynamically conditioning the shared expert on the translation direction via FiLM modulation (Perez et al., 2018),

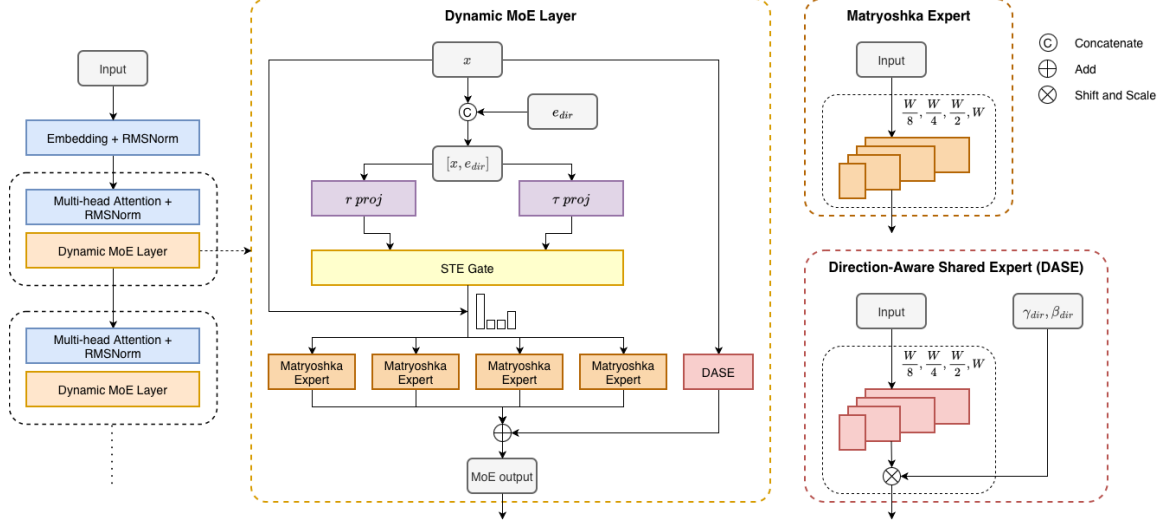


Figure 1: Overview of the **EDWARD-NMT** architecture. **Left:** The macro-level Transformer block integrating the Dynamic MoE layer. **Center:** The Adaptive Sigmoid Router computes a token-specific threshold (τ) conditioned on a translation direction embedding (e_{dir}) to dynamically allocate the sparse expert budget. **Top Right:** The internal structure of a Matryoshka Expert, illustrating the nested parameter slices used to extract varying capacities. **Bottom Right:** The Direction-Aware Shared Expert (DASE) applies FiLM modulation ($\gamma_{dir}, \beta_{dir}$) to provide a stable, direction-aware backbone.

creating a unified, direction-aware backbone without the parameter overhead of dedicated experts.

3 EDWARD-NMT

EDWARD-NMT builds upon the Transformer encoder-decoder architecture (Vaswani et al., 2017; Shazeer, 2020), selected over decoder-only alternatives for its superior parameter efficiency and low inference latency, making it highly practical for resource-constrained deployment. We replace conventional feed-forward sublayers with a novel MoE block comprising three jointly optimized innovations: (i) **elastic experts** encoding a family of nested inner widths, (ii) an **adaptive sigmoid router** dynamically allocating the optimal expert count per token, and (iii) a **dual-consistency** objective aligning predictions across capacity views. Additionally, a lightweight **Direction-Aware Shared Expert (DASE)** is computed in parallel. Figure 1 provides a schematic overview.

3.1 Preliminaries: Standard Mixture-of-Experts

We adopt the sparsely gated MoE feed-forward layer (Shazeer et al., 2017; Fedus et al., 2022) as our baseline. Let $x \in \mathbb{R}^{d_{\text{model}}}$ denote the input token representation. An MoE layer consists of E independent expert networks, $\{E_i\}_{i=1}^E$, and a routing function $g_{\theta,k}(x)$ parameterized by a weight

matrix $W_g \in \mathbb{R}^{d_{\text{model}} \times E}$.

The standard router projects the input to unnormalized logits $z = xW_g$. The routing weights are thus computed as $g_{\theta,k}(x) = \text{Softmax}(\text{TopK}(xW_g, k))$, and the MoE layer outputs their weighted sum:

$$\text{MoE}(x) = \sum_{i \in \mathcal{T}_k(x)} g_{\theta,k}(x)_i \cdot E_i(x), \quad (1)$$

where $\mathcal{T}_k(x)$ is the set of selected expert indices. Each expert E_i is typically a SwiGLU Multi-Layer Perceptron (Shazeer, 2020), parameterized by up-projection weights $W_1, W_2 \in \mathbb{R}^{d_{\text{model}} \times M}$ and down-projection weights $W_3 \in \mathbb{R}^{M \times d_{\text{model}}}$, where M is the maximum inner dimension.

This standard formulation exhibits two critical limitations: rigid top- k routing fixes the per-token compute budget, and the physical memory footprint scales linearly with $E \cdot M$ regardless of how many experts are actively evaluated.

3.2 Elastic Experts via Matryoshka Dimensions

Inspired by Matryoshka Representation Learning (Kusupati et al., 2022; Kudugunta et al., 2024), we design experts to support a nested family of inner dimensions $\mathcal{D} \subseteq \{1, \dots, M\}$ sharing a *single* set of parameter tensors. During each forward pass, a target dimension $D_k \in \mathcal{D}$ is sampled. Rather

than instantiating physically separate experts, we dynamically truncate intermediate activations to their first D_k components. Letting $[:, D_k]$ denote the hidden dimension slice, the elastic SwiGLU expert computes:

$$h = \text{SiLU}(xW_1) \odot (xW_2), \quad (2)$$

$$E_i(x; D_k) = \psi(D_k) \cdot (h[:, D_k] W_{3, [:D_k, :]}) \cdot (3)$$

Variance Preservation. Truncating the inner dimension intrinsically shrinks the magnitude of the down-projected activations, destabilizing residual additions and RMSNorm statistics (Zhang and Senrich, 2019). To compensate, we introduce a deterministic rescaling factor $\psi(D_k)$ dependent strictly on the chosen width:

$$\psi(D_k) = \sqrt{\frac{M}{\max(D_k, 1)}}. \quad (4)$$

This ensures the expected squared norm of $E_i(x; D_k)$ remains invariant to D_k , a correction paramount for stable joint training across all nested capacities. Consequently, a single trained network seamlessly exposes $|D|$ weight-sharing sub-networks freely selectable at inference.

3.3 Adaptive Sigmoid Routing

To decouple the model from rigid top- k constraints, we replace the standard softmax gate with a token-adaptive *sigmoid* gate, empowering tokens to activate a variable expert count based on complexity.

Differentiable τ -routing. Let $z = xW_g \in \mathbb{R}^E$ be the unnormalized router logits with learned projection $W_g \in \mathbb{R}^{d_{\text{model}} \times E}$, and $e_{\text{dir}} \in \mathbb{R}^{d_{\text{dir}}}$ be a learnable direction embedding for the translation direction. We compute per-expert probabilities r and a dynamic threshold τ :

$$r = \sigma(z + \epsilon), \quad \epsilon \sim \mathcal{U}(-\eta, \eta), \quad (5)$$

$$\tilde{\tau} = \sigma([x; e_{\text{dir}}]^\top W_\tau), \quad (6)$$

$$\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \tilde{\tau}, \quad (7)$$

where $W_\tau \in \mathbb{R}^{d_{\text{model}} + d_{\text{dir}}}$ projects the concatenated representation, and η is a training jitter to prevent routing collapse (Fedus et al., 2022).

A token activates expert E_i if and only if $r_i > \tau$. To backpropagate through this discrete decision, we use a Straight-Through Estimator (STE) (Bengio et al., 2013):

$$M_i^{\text{soft}} = \sigma(r_i - \tau), \quad (8)$$

$$M_i^{\text{hard}} = \mathbb{I}[r_i > \tau], \quad (9)$$

$$M_i = \text{sg}(M_i^{\text{hard}} - M_i^{\text{soft}}) + M_i^{\text{soft}}, \quad (10)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. The MoE output is given by:

$$\text{MoE}(x) = \sum_{i=1}^E M_i r_i E_i(x; D_k). \quad (11)$$

During inference, only experts where $M_i^{\text{hard}} = 1$ are activated, rendering per-token compute dynamically data-dependent.

Expert Sparsity Constraints. To explicitly control the computational budget and prevent load imbalance, we enforce a target sparsity ρ_{target} for active experts. An auxiliary loss aligns the empirical routing sparsity with this constraint to ensure uniform expert utilization:

$$\begin{aligned} \mathcal{L}_{\text{sp}} = & \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left(\sum_{i=1}^E M_i(x) - \rho_{\text{target}} \right)^2 \\ & + \frac{1}{E} \sum_{i=1}^E \left(f_i - \frac{\rho_{\text{target}}}{E} \right)^2. \end{aligned} \quad (12)$$

where f_i is the activation frequency of expert i within batch $\mathcal{B}$, masking padded positions for accurate sparsity estimation.

3.4 Dual-Consistency Objective

A central challenge in elastic MoEs is ensuring aggressively truncated sub-networks remain performant without downstream fine-tuning. We resolve this via a *dual-consistency* objective tying varying capacity views together.

At every training step, we sample two dimensions $D_A, D_B \sim \mathcal{D}$ with replacement. We perform two forward passes by dynamically slicing expert activations to widths D_A and D_B , yielding predictive vocabulary distributions P_A and P_B . We define the loss using a Symmetrized Kullback-Leibler (KL) divergence (Jeffreys Divergence):

$$\mathcal{L}_{\text{dc}} = \frac{1}{2} \left[\text{KL}(\text{sg}(P_A) \| P_B) + \text{KL}(\text{sg}(P_B) \| P_A) \right], \quad (13)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This formulation unifies two regularizing paradigms. When $D_A \neq D_B$, it functions as *in-place self-distillation*: the larger capacity view provides a richer target to guide the smaller variant, while the smaller core acts as a structural regularizer for the larger network. When $D_A = D_B$, it defaults to consistency regularization similar to

R-Drop (Liang et al., 2021), minimizing the divergence between stochastic forward passes of the same sub-network. This bidirectional flow anchors the full-capacity model while preserving the translation quality of truncated variants.

3.5 Direction-Aware Shared Expert

Multi-directional and multi-domain NMT models frequently suffer from negative interference. Following recent successes isolating common knowledge in shared experts (Rajbhandari et al., 2022a; Dai et al., 2024), we introduce a single *dense* expert activated for every token in parallel with the sparse router. To render this backbone direction-aware without the parameter overhead of dedicated experts, we dynamically modulate its hidden state via Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). We use the discrete translation direction to index learned modulation vectors $\gamma_c, \beta_c \in \mathbb{R}^M$. The DASE computes:

$$h^{\text{shr}} = \text{SiLU}(xW_1^{\text{shr}}) \odot (xW_2^{\text{shr}}), \quad (14)$$

$$\hat{h}^{\text{shr}} = h^{\text{shr}} \odot (1 + \gamma_c) + \beta_c, \quad (15)$$

$$E_{\text{shr}}(x; D_k) = \alpha \cdot \psi(D_k) \cdot (\hat{h}_{[:,D_k]}^{\text{shr}} W_{3,[:,D_k,:]}^{\text{shr}}), \quad (16)$$

where $0 < \alpha < 1$ is a static damping factor preventing the shared expert from overpowering the sparsely routed experts. The DASE acts as a stable, grammatical backbone, freeing routed experts to specialize heavily across idiosyncratic domains.

The final output of the elastic MoE block aggregates the dynamic routed paths with the shared direction-conditioned path:

$$y_{\text{MoE}}(x; D_k) = \sum_{i=1}^E M_i r_i E_i(x; D_k) + E_{\text{shr}}(x; D_k). \quad (17)$$

3.6 Training Objective and Inference

The complete optimization objective minimizes the standard label-smoothed cross-entropy loss alongside our auxiliary routing constraints and dual-consistency regularization. Because we sample two distinct capacity dimensions D_A and D_B at each training step, the cross-entropy and routing losses are calculated independently for each width and averaged. The final loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \frac{1}{2} \sum_{D \in \{D_A, D_B\}} \left(\mathcal{L}_{\text{ce}}^{(D)} + \lambda_{\text{sp}} \mathcal{L}_{\text{sp}}^{(D)} \right) + \lambda_{\text{dc}} \mathcal{L}_{\text{dc}}, \quad (18)$$

where $\mathcal{L}_{\text{ce}}^{(D)}$ and $\mathcal{L}_{\text{sp}}^{(D)}$ denote the cross-entropy and target sparsity computed during the forward pass of width D . The scalar weights λ_{sp} and λ_{dc} balance the relative impact of the routing and consistency constraints.

During inference, a target memory footprint is dictated simply by specifying an expert width $D_k \in \mathcal{D}$. The corresponding sub-network is extracted seamlessly via deterministic matrix slicing. Simultaneously, the adaptive sigmoid router dynamically calibrates the active expert count per token. By decoupling static memory capacity from the dynamic compute budget, EDWARD-NMT realizes a highly efficient “train once, deploy anywhere” paradigm.

4 Experimental Setup

4.1 Datasets

We evaluate EDWARD-NMT on four benchmark families (full dataset and preprocessing details are provided in Appendix A.1): (1) **PhoMT** (En $\leftrightarrow$ Vi) (Doan et al., 2021), (2) the standard **WMT-14** (En $\leftrightarrow$ De) (Bojar et al., 2014) evaluated on newstest2014, (3) the **multi-domain** De $\leftrightarrow$ En benchmark (Zhang et al., 2024) spanning five diverse domains evaluated via macro-averaging, and (4) the multilingual translation **OPUS-16** dataset following Zhao et al. (2024), with 16 languages divided into high ($> 0.9\text{M}$ pairs), low ($< 0.1\text{M}$ pairs) and medium resource.

4.2 Architecture and Sub-network Family

The encoder-decoder backbone has $L_{\text{enc}} = L_{\text{dec}} = 8$ layers, $d_{\text{model}} = 512$, eight attention heads, $E = 4$ routed experts, and one direction-aware shared expert (total of 5 experts) per MoE layer. Both routed and shared experts have a maximum inner dimension $M = 2048$, governed by the nested family: $\mathcal{D} = \{2048, 1024, 512, 256\}$. This yields the four variants in Table 1. Total parameters vary across datasets due to vocabulary size. All share *a single set of trained parameters*, differing only by the inference slice $D_k \in \mathcal{D}$.

Variant	D_k	Total Params	Active Params
EDWARD-NMT-S	256	67 – 74M	51 – 59M
EDWARD-NMT-M	512	98 – 105M	66 – 76M
EDWARD-NMT-L	1024	161 – 168M	97 – 109M
EDWARD-NMT-XL	2048	287 – 294M	158 – 175M

Table 1: EDWARD-NMT sub-network variants extracted from a single trained model.

Routing & Consistency. The direction embedding $e_{\text{dir}} \in \mathbb{R}^{64}$ is learned per translation direction. The router uses jitter $\eta = 0.05$, thresholds $\tau_{\text{min}} = 0.05$ and $\tau_{\text{max}} = 0.60$. The target sparsity is set to $\rho_{\text{target}} = 1.25$. The auxiliary weight is $\lambda_{\text{sp}} = 0.01$, and the DASE damping factor is $\alpha = 0.1$. For dual-consistency ($\lambda_{\text{dc}} = 1$), dimensions $D_A, D_B \sim \mathcal{D}$ are sampled uniformly with replacement per step.

4.3 Baselines

We compare against standard dense Transformers (BASE/BIG) (Vaswani et al., 2017; Ott et al., 2018), mBART (Liu et al., 2020), and MoE methods: the Switch Transformer (Fedus et al., 2022) with top-1 token-based routing (ST-MoE) and GShard (Lepikhin et al., 2021) with top-2 token-based routing (GS-MoE). Both MoE models use TRANSFORMERS BIG backbone for multi-domain benchmark. All models are *trained jointly on all translation directions* to match our setup. For OPUS-16 we report baseline results from Zhao et al. (2024), more MoE baselines are provided in Appendix B.2.

4.4 Implementation Details

Training. All models are trained on a single TPU v5-64 pod using the AdamW optimizer (Loshchilov and Hutter, 2019) ($\beta_1=0.9$, $\beta_2=0.98$). The learning rate warms up linearly from 10^{-6} to 10^{-2} over the first 1k steps. We use an effective batch size of 8,192 sequences (max length 128), a dropout rate of 0.1, and label smoothing of 0.1. Training spans 16k steps for the multi-domain, 30k for OPUS-16, 40k for PhoMT, and 60k for WMT-14. Complete hyperparameters are listed in Appendix Table 10.

Inference and Evaluation. We decode via beam search (width 4, length penalty 0.6). Performance is measured using case-sensitive sacreBLEU (Post, 2018) for PhoMT, WMT-14, and OPUS-16, and tokenized BLEU (Papineni et al., 2002) for the multi-domain dataset. We also report COMET (Rei et al., 2020) for semantic fidelity on WMT-14 and multi-domain dataset. chrF2 (Popović, 2016) on WMT-14 and TER (Snover et al., 2006) on PhoMT results are detailed in Appendix B.6 and B.7. All reported improvements over baselines are statistically significant (Appendix A.4). Edge inference latency and throughput are profiled using compiled LiteRT checkpoints (Appendix B.1). Crucially, all four EDWARD-NMT variants are extracted and evaluated from the *same* checkpoint.

5 Results

5.1 Main Translation Quality

PhoMT and WMT-14. Table 2 reports sacreBLEU on PhoMT (En $\leftrightarrow$ Vi) and WMT-14 (En $\leftrightarrow$ De). EDWARD-NMT-XL exceeds TRANSFORMER BIG across all directions while activating fewer parameters, improving by +1.30 BLEU on PhoMT (Vi $\rightarrow$ En) and +1.48 on WMT-14 (De $\rightarrow$ En). Furthermore, it surpasses much larger mBART models across all directions on En $\rightarrow$ Vi, while remaining highly competitive on Vi $\rightarrow$ En despite using $3\times$ fewer active parameters. The smaller sub-networks extracted from the same checkpoint maintain robust performance: EDWARD-NMT-S (51 – 59M active parameters) outperforms the dense TRANSFORMER BASE by +0.69 to +3.18 BLEU at a comparable parameter footprint, performing on par with TRANSFORMER BIG on PhoMT.

Multi-domain De $\leftrightarrow$ En. Table 3 compares EDWARD-NMT to dense and MoE baselines on the multi-domain benchmark. Our EDWARD-NMT-XL achieves the highest average BLEU, outperforming TRANSFORMER BIG by +0.62 BLEU across all five domains with significantly fewer active parameters (161M vs. 210M). This efficiency extends to smaller sub-networks: EDWARD-NMT-L surpasses TRANSFORMER BIG using under half the active parameters (102M), while the smallest EDWARD-NMT-S (57M) improves upon TRANSFORMER BASE by +0.49 BLEU and rivals the much heavier GS-MoE (260M).

Multilingual OPUS-16. Demonstrating robust scalability, our architecture also excels on the massively multilingual OPUS-16 dataset (Table 3). While our most compressed sub-network, EDWARD-NMT-S, remains competitive, scaling to just the EDWARD-NMT-M variant is sufficient to consistently outperform both ST-MoE and GS-MoE across both translation directions. The full-capacity EDWARD-NMT-XL further extends this lead. An extended comparison against a wider range of baselines is provided in Appendix B.2.

Semantic Quality Beyond surface-level n -gram matching, we evaluate semantic integrity using COMET-22¹ (Rei et al., 2022). Table 4 presents the average scores across all evaluation directions

¹<https://huggingface.co/Unbabel/wmt22-comet-da>

Method	PhoMT				WMT-14			
	Params	Active	En→Vi	Vi→En	Params	Active	En→De	De→En
TRANSFORMER BASE	65M	65M	42.12	37.19	61M	61M	24.56	28.76
TRANSFORMER BIG	213M	213M	42.94	37.83	210M	210M	27.53	31.48
mBART [†]	448M	448M	43.46	39.78	610M	610M	24.90	32.51
EDWARD-NMT-S	67M	51M	42.91	37.88	74M	59M	26.85	31.94
EDWARD-NMT-M	98M	66M	43.35	38.50	105M	76M	27.49	32.45
EDWARD-NMT-L	161M	97M	<u>43.61</u>	38.93	168M	109M	<u>27.96</u>	<u>32.65</u>
EDWARD-NMT-XL	287M	158M	43.79	<u>39.13</u>	294M	175M	28.03	32.96

Table 2: Main results on PhoMT (En↔Vi) and WMT-14 (En↔De) measured in sacreBLEU. “Active” denotes average parameters activated by our sigmoid router. Best and second-best scores are **bolded** and underlined, respectively. [†]For mBART on WMT-14, we use a pretrained many-to-many model (mbart-large-50-many-to-many-mmt).

Method	Params	Active	Multi-domain						OPUS-16	
			KOR	IT	MED	LAW	SUB	AVG	En→XX	XX→En
TRANSFORMER BASE	61M	61M	20.53	40.01	51.25	53.72	25.29	38.16	26.92 [†]	30.67 [†]
TRANSFORMER BIG	210M	210M	22.55	<u>41.84</u>	<u>53.35</u>	56.22	26.73	40.13	-	-
ST-MoE	360M	210M	21.73	41.00	52.01	55.69	26.18	39.32	29.68 [†]	33.89 [†]
GS-MoE	360M	260M	20.70	41.35	52.31	54.56	24.72	38.72	28.28 [†]	32.31 [†]
EDWARD-NMT-S	74M	57M	21.61	40.07	51.01	53.99	26.56	38.65	29.55	32.99
EDWARD-NMT-M	105M	72M	22.72	40.87	52.11	55.21	26.88	39.56	30.73	34.84
EDWARD-NMT-L	168M	102M	<u>23.56</u>	41.53	52.95	<u>56.27</u>	<u>27.13</u>	40.28	<u>31.64</u>	<u>35.82</u>
EDWARD-NMT-XL	294M	161M	24.06	41.90	53.45	57.05	27.30	40.75	32.27	36.30

Table 3: Experimental results on the multi-domain (De↔En) averaged across both directions and multilingual OPUS-16 benchmarks (English to any En→XX and any to English XX→En). The “Params” and “Active” columns report footprints for the multi-domain setup. [†]Reported from Zhao et al. (2024). Extended OPUS-16 comparisons are provided in Appendix B.2.

Method	WMT-14		Multi-domain	
	Active	AVG	Active	AVG
TRANSFORMER BASE	61M	80.78	61M	80.20
TRANSFORMER BIG	210M	83.69	210M	<u>81.62</u>
mBART [†]	610M	83.56	-	-
ST-MoE	-	-	210M	80.95
GS-MoE	-	-	260M	78.20
EDWARD-NMT-S	59M	83.48	57M	80.96
EDWARD-NMT-M	76M	83.98	72M	81.32
EDWARD-NMT-L	109M	84.12	102M	81.62
EDWARD-NMT-XL	175M	84.30	161M	81.80

Table 4: Semantic-level translation quality measured in COMET-22. Scores are averaged across all translation directions (and domains for the Multi-domain benchmark). Best and second-best scores are **bolded** and underlined. [†]mBART is a pretrained many-to-many model.

and domains. On WMT-14, our highly compressed EDWARD-NMT-S (59M active parameters) significantly outperforms TRANSFORMER BASE (61M active) by an average margin of +2.70 COMET. Additionally, our EDWARD-NMT-M (76M) and EDWARD-NMT-L (109M) variants

consistently exceed the 210M-parameter TRANSFORMER BIG. At full capacity, EDWARD-NMT-XL establishes a new Pareto frontier by surpassing TRANSFORMER BIG by +0.61 COMET on average. This semantic robustness extends to the Multi-domain benchmark, where heavily parameterized models such as GS-MoE (260M) experience severe semantic degradation (78.20 COMET). In stark contrast, our EDWARD-NMT-S model (57M) maintains a robust score of 80.96 COMET. This performance is comparable to the 210M-parameter ST-MoE and outperforms GS-MoE by +2.76 points. Furthermore, the EDWARD-NMT-XL model achieves the highest overall score (81.80), demonstrating that our architecture strictly preserves meaning even within aggressively truncated sub-networks.

Quality vs. Active Parameters. Driven by our adaptive sigmoid routing, which activates only 1.34 to 1.70 experts per token (Table 5), EDWARD-NMT establishes an efficient Pareto frontier. Crucially, this number of active experts is explicitly controlled by the target density ρ (Table 25).

Method	PhoMT	WMT-14	MD	OPUS-16
EDWARD-NMT-S	1.44	1.64	1.37	1.34
EDWARD-NMT-M	1.45	1.66	1.37	1.35
EDWARD-NMT-L	1.47	1.69	1.38	1.37
EDWARD-NMT-XL	1.51	1.70	1.39	1.39

Table 5: Average active **routed experts** per token across the PhoMT, WMT-14, Multi-domain (MD) and OPUS-16 benchmarks (out of $E = 4$ routed experts, excluding the fixed shared expert).

The extracted sub-networks consistently outperform dense baselines like TRANSFORMER BIG and mBART while using up to $3.6\times$ fewer active parameters. For instance, EDWARD-NMT-L surpasses TRANSFORMER BIG utilizing less than half the active capacity (102M vs. 210M) and higher generation throughput (61.5 vs. 48.0 TPS). This structural compactness translates effectively to practical on-device acceleration. While experiencing only a marginal drop in translation quality, the compact S variant reduces the total parameter footprint by $4\times$ and delivers a $2.2\times$ speedup over the full XL variant (Appendix Table 11), with all sizes extracted from a single checkpoint.

5.2 Ablation Studies

All ablations are run on the multi-domain benchmark (averaged across En $\leftrightarrow$ De) and report the tokenized BLEU of every sub-network (XL/L/M/S) extracted from the same training run. PhoMT ablations are provided in Appendix B.7.

Contributions of Each Component. Table 6 incrementally adds EDWARD-NMT components to a Top-2 MoE baseline. Adding Matryoshka Experts (ME) yields immediate gains, peaking the XL capacity at 40.96. While fully integrating the Sigmoid router and Dual-Consistency (DC) causes a slight XL quality drop on this benchmark, this represents a deliberate, highly favorable trade-off. The adaptive router drastically reduces the active compute budget (Top-2 $\rightarrow \sim 1.38$ experts), while DC prevents smaller sub-networks from collapsing. Consequently, the full system (+ ME + DC + Sigmoid) provides the optimal configuration, maintaining highly competitive XL performance at significantly lower inference costs, while maximizing translation quality for the L, M, and S variants. Ablations on PhoMT (Appendix Table 20) confirm the efficacy that it improves all capacities performance.

Setting	En $\leftrightarrow$ De			
	XL	L	M	S
MoE ($k=2$)	40.35	39.64	39.21	38.24
+ ME	40.96	<u>40.26</u>	39.35	38.34
+ ME + Sigmoid	40.72	40.23	39.20	38.16
+ ME + DC	40.66	40.16	<u>39.36</u>	<u>38.49</u>
+ ME + DC + Sigmoid	<u>40.75</u>	40.28	39.56	38.65

Table 6: Ablation results on the Multi-domain benchmark, averaged over both translation directions. *All settings use shared expert (DASE).*

Setting	En $\leftrightarrow$ De			
	XL	L	M	S
No SE	39.75	39.40	38.84	38.13
Common SE	<u>39.95</u>	<u>39.73</u>	<u>39.12</u>	38.46
DASE (ours)	40.74	40.26	39.39	38.72

Table 7: Ablation on the shared expert (SE) routing strategies on the Multi-domain benchmark, averaged over both translation directions.

Setting	En $\leftrightarrow$ De			
	XL	L	M	S
No loss	40.72	<u>40.23</u>	39.20	38.16
R-Drop	40.00	39.73	39.19	38.47
Sym. JSD	<u>40.80</u>	40.17	39.19	38.19
Asym. KL	40.76	40.14	<u>39.50</u>	<u>38.59</u>
Asym. KL (Max)	41.04	39.89	39.14	38.39
DC Loss (ours)	40.75	40.28	39.56	38.65

Table 8: Ablation on the Dual-Consistency loss (DC Loss) on the Multi-domain benchmark, averaged over both translation directions.

Impact of Shared Expert Routing. Table 7 isolates the effect of the shared expert strategy. Compared to a standard MoE without shared experts (No SE), introducing a Common Shared Expert (Common SE) provides modest gains (+0.2 to +0.3 BLEU). However, our Direction-Aware Shared Expert (DASE) unlocks substantial improvements across the board, yielding approximately +1.0 BLEU on XL and +0.6 BLEU on S over the No SE baseline. This confirms that intelligently routing common knowledge is vital for maximizing multi-domain translation quality.

Choice of Consistency Loss. Our proposed Dual-Consistency (**DC Loss**) introduces a bidirectional alignment mechanism critical for elastic architectures. As shown in Table 8, standard R-Drop isolates identical widths, which limits cross-scale knowledge transfer and degrades the flagship capacity (−0.6 BLEU on L). A symmetric Jensen-

Shannon Divergence (Sym. JSD) (Lin, 1991) is similarly ineffective, providing negligible gains for smaller sub-networks. While unidirectional distillation (Asym. KL) mitigates this by treating the larger sampled width as a detached teacher, it lacks mutual regularization. Furthermore, anchoring strictly to the full-capacity model (Asym. KL Max) yields sub-optimal intermediate variants. By coupling capacity views - with larger variants serving as distillation targets and smaller cores acting as regularizers - our DC Loss establishes a balance that maximizes quality across all sub-networks.

Elasticity of Model Capacities. EDWARD-NMT’s structural elasticity enables fluid extraction of sub-networks at arbitrary granularities. As detailed in Table 14 (Appendix), despite being optimized on only four discrete views ($\mathcal{D} = \{2048, 1024, 512, 256\}$), the capacity space seamlessly interpolates. Unseen intermediate sub-networks (e.g., $D = 1536$ or 768) maintain robust translation quality and strictly follow a monotonic compute-performance trade-off without post-hoc fine-tuning. However, this elasticity is bounded by the trained envelope; extrapolating below the minimum trained threshold ($D = 128$) triggers a severe performance breakdown, confirming that structural integrity requires optimized geometric bounds.

Effect of Variance Preservation. Removing the deterministic rescaling factor $\psi(D_k)$ consistently degrades translation quality across all extracted sub-networks (Table 15, Appendix). This performance drop becomes significantly more pronounced at smaller capacities (-0.37 BLEU for $D = 256$), confirming that explicitly stabilizing down-projected activation magnitudes is crucial for the robust joint training of nested structures.

6 Conclusion

We introduced **EDWARD-NMT**, an elastic MoE architecture realizing an efficient *train-once, deploy-anywhere* paradigm for NMT. By synergizing Matryoshka-style elastic experts, an adaptive sigmoid router, and a direction-aware shared expert, our framework decouples the static memory footprint from the dynamic computational budget. Crucially, a symmetric dual-consistency objective aligns predictive distributions across nested widths, ensuring truncated sub-networks retain high quality without post-hoc fine-tuning. Extensive evaluations across PhoMT, WMT-14, multilingual

OPUS-16, and a multi-domain benchmark validate our approach. The full-capacity EDWARD-NMT matches or exceeds strong MoE baselines, while the seamlessly extracted smaller sub-networks remain highly competitive, outperforming comparably sized dense Transformers. Overall, coupling structural elasticity with dynamic routing offers a practical, scalable path to resource-aware sequence modeling and adaptable on-device deployment.

Limitations

Despite its effectiveness, EDWARD-NMT has several limitations that offer avenues for future research:

Architectural Scope. We evaluate EDWARD-NMT strictly on encoder-decoder architectures to demonstrate its efficacy for machine translation targeted at resource-constrained edge devices. Extending this elastic formulation to massively multilingual settings, decoder-only large language models, or broader sequence generation tasks remains an important direction for future work.

Dimensions of Elasticity. Our design prioritizes structural flexibility within the feed-forward networks, as they constitute the vast majority of parameters in MoE models. Although EDWARD-NMT achieves robust elasticity for these capacities within the optimized training envelope, we have not yet explored elasticity along orthogonal architectural axes. Future work will investigate integrating early exiting (dynamic depth) and variable hidden state dimensions (d_{model}) to realize a fully elastic, multi-dimensional architecture capable of simultaneous depth-and-width adaptation.

Extreme Compression. Although EDWARD-NMT interpolates robustly within its trained envelope ($\mathcal{D}$), performance inherently degrades when extrapolating below the minimum trained capacity. Investigating continuous width formulations or dynamically expanding these lower bounds without destabilizing the shared backbone remains a key direction for extreme-compression scenarios.

Ethical Considerations

This work uses only publicly available NMT benchmarks with no personally identifiable or sensitive data, and all baselines are cited and reproduced under their respective licenses. Like any NMT system, EDWARD-NMT may propagate biases or inaccuracies from training corpora and should not inform

high-stakes decisions without human review. We foresee no direct harmful applications; by improving parameter and compute efficiency, our work lowers the barrier to deploying translation systems on resource-constrained devices, improving accessibility to translation technology.

Acknowledgments

This research was supported with Cloud TPUs from Google’s TPU Research Cloud (TRC) program. We would like to thank the TRC team for providing the computational resources used in this work.

References

- Roei Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7747–7763.
- Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.
- Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the second conference on machine translation*, pages 118–126.
- Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2020. Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*.
- Umberto Cappellazzo, Minsu Kim, Pingchuan Ma, Honglie Chen, Xubo Liu, Stavros Petridis, and Maja Pantic. 2026. Mome: Mixture of matryoshka experts for audio-visual speech recognition. *Advances in neural information processing systems*, 38:40013–40040.
- Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, and 1 others. 2022. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.
- Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Stablemoe: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7085–7095.
- Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4495–4503.
- Maha Elbayad, Anna Sun, and Shruti Bhosale. 2023. [Fixing MoE over-fitting on low-resource languages in multilingual machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14237–14253, Toronto, Canada. Association for Computational Linguistics.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Tommaso Furlanello, Zachary Lipton, Michael Tschanen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR.
- Baban Gain, Dibyanayan Bandyopadhyay, Asif Ekbal, and Trilok Nath Singh. 2026. Bridging the linguistic divide: a survey on leveraging large language models for machine translation. *Language Resources and Evaluation*, 60(2):40.
- Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, Zhaopeng Tu, and Tao Lin. 2025. Dynamic mixture of experts: An auto-tuning approach for efficient transformer models. In *International Conference on Learning Representations*, volume 2025, pages 79643–79672.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030.

- Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.
- Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks VS Lakshmanan, Ahmed Hassan Awadallah, Sebastien Bubeck, and Jianfeng Gao. 2023. Auto-moe: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9116–9132.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open source toolkit for statistical machine translation](#). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the first workshop on neural machine translation*, pages 28–39.
- 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, Brussels, Belgium. Association for Computational Linguistics.
- Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. [Beyond distillation: Task-level mixture-of-experts for efficient inference](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3577–3599, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hananeh Hajishirzi, Sham Kakade, Ali Farhadi, and Prateek Jain. 2024. Matformer: Nested transformer for elastic inference. *Advances in Neural Information Processing Systems*, 37:140535–140564.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.
- François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629.
- Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Knowledge distillation by on-the-fly native ensemble. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7528–7538.
- Dmitry Lepikhin, HyounJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. [GShard: Scaling giant models with conditional computation and automatic sharding](#). In *ICLR*.
- Zichong Li, Chen Liang, Zixuan Zhang, Ilgee Hong, Young Jin Kim, Weizhu Chen, and Tuo Zhao. 2025. [Slimmoe: Structured compression of large moe models via expert slimming and distillation](#). In *Second Conference on Language Modeling*.
- Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*, 34:10890–10905.
- Yunlong Liang, Fandong Meng, and Jie Zhou. 2025. Thor-moe: Hierarchical task-guided and context-responsive routing for neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21433–21445.
- J. Lin. 1991. [Divergence measures based on the Shannon entropy](#). *IEEE Transactions on Information Theory*, 37(1):145–151.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.
- Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. 2020. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361.

- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.
- 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- MinhQuang Pham, Josep Maria Crego, and François Yvon. 2021. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9:17–35.
- Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. [Learning language-specific layers for multilingual machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14767–14783, Toronto, Canada. Association for Computational Linguistics.
- Maja Popović. 2016. [chrF deconstructed: beta parameters and n-gram weights](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. 2024. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pages 28435–28445.
- Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022a. DeepSpeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pages 18332–18346. PMLR.
- Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022b. [DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18332–18346. PMLR.
- Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- 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, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.
- Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.
- 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. In *International Conference on Learning Representations*.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Felix Stahlberg. 2020. [Neural machine translation: A review](#). *Journal of Artificial Intelligence Research*, 69:343–418.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [Roformer: Enhanced transformer with rotary position embedding](#). *Neurocomputing*, 568(C).
- Nurbek Tastan, Stefanos Laskaridis, Karthik Nandakumar, and Samuel Horváth. 2026. [MoSE: Mixture of slimmable experts for efficient and adaptive language models](#). In *Forty-third International Conference on Machine Learning*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ziteng Wang, Jun Zhu, and Jianfei Chen. 2025. Remoe: Fully differentiable mixture-of-experts with relu routing. In *International Conference on Learning Representations*, volume 2025, pages 59486–59507.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Haoran Xu, Maha Elbayad, Kenton Murray, Jean Mailard, and Vedanuj Goswami. 2023. Towards being parameter-efficient: a stratified sparsely activated transformer with dynamic capacity. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12858–12870.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zengwei Yao, Wei Kang, Xiaoyu Yang, Fangjun Kuang, Liyong Guo, Han Zhu, Zengrui Jin, Zhaoqing Li, Long Lin, and Daniel Povey. 2025. Cr-ctc: Consistency regularization on ctc for improved speech recognition. In *International Conference on Learning Representations*, volume 2025, pages 26850–26868.

Jiahui Yu and Thomas S Huang. 2019. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1803–1811.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. 2019. Slimmable neural networks. In *7th International Conference on Learning Representations, ICLR 2019*.

Zihao Zeng, Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. 2024. Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6223–6235.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in neural information processing systems*, 32.

Fan Zhang, Mei Tu, Song Liu, and Jinyao Yan. 2024. A lightweight mixture-of-experts neural machine translation model with stage-wise training strategy. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2381–2392.

Xinyu Zhao, Xuxi Chen, Yu Cheng, and Tianlong Chen. 2024. Sparse moe with language guided routing for multilingual machine translation. In *The Twelfth International Conference on Learning Representations*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y Zhao, Andrew M Dai, Zhifeng Chen, Quoc V Le, and James Laudon. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

A Experimental Details

This appendix provides the comprehensive experimental details required to reproduce the results reported in Section 5. We outline the datasets, training configurations, and hyperparameters necessary to independently re-implement the methodology (Section 3) using publicly available data.

A.1 Datasets and Preprocessing

The datasets are summarized in Table 9.

Dataset	Lang	Train	Dev	Test	Vocab
PhoMT	En↔Vi	2,977k	19,151	19,151	20k SPM
WMT-14	En↔De	3,906k	3,000	3,003	32k BPE
MD	De↔En	1,284k	10,000	10,000	32k BPE
OPUS-16	16 lang	17,559k	30,000	30,000	32k SPM

Table 9: Dataset sizes (sentence pairs) and shared vocabularies. SPM: SentencePiece. The “Train” column reports the final number of filtered sentence pairs used for model training. The multi-domain (MD) training set is the concatenation of five diverse domains (Koran, IT, Medical, Law, Subtitles) following Zhang et al. (2024). The OPUS-16 dataset includes 16 languages following Zhao et al. (2024).

PhoMT (En↔Vi). A large-scale parallel corpus (Doan et al., 2021) comprising approximately 3M sentence pairs across domains such as news, subtitles, legal documents, and Wikipedia. We use the full PhoMT corpus, utilizing the official training pairs directly. For subword tokenization, we train a shared SentencePiece (Kudo and Richardson, 2018) model with a vocabulary size of 20,000 tokens strictly on the training set, which is subsequently applied to encode all data splits.

WMT-14 (En↔De). We utilize the standard WMT-14 training corpora (Europarl v7, Common Crawl, and News Commentary v9), comprising approximately 4.5M original sentence pairs. First, we normalize punctuation and tokenize the text using the Moses toolkit (Koehn et al., 2007)². Next,

²<https://github.com/moses-smt/mosesdecoder>

we learn a joint Byte-Pair Encoding (BPE) (Sennrich et al., 2016) model with 32,000 merge operations on the concatenated English and German training data, which is subsequently applied to all data splits. Finally, we filter the training and validation sets to remove sentences longer than 250 tokens or with a source-target length ratio exceeding 1.5. This results in roughly 3.9M filtered training pairs. Final translation performance is evaluated on the standard newstest2014 test set. The models are trained using a joined dictionary capped at 32,768 tokens. Our preprocessing pipeline is based on open-source Fairseq³ (Ott et al., 2019).

Multi-domain De↔En. Following Zhang et al. (2024), we evaluate on a multi-domain corpus encompassing five distinct domains: KORAN, IT, MEDICAL, LAW, and SUBTITLES. The total filtered size is approximately 1.3M sentence pairs. The tokenization and normalization process follows the exact same Moses-based pipeline as WMT-14. We aggregate the training sets from all five domains across both languages to learn a single, unified BPE model with 32,000 merge operations. After applying the identical length and ratio filtering (maximum length of 250, length ratio of 1.5), the filtered data from all domains is concatenated into a single combined training set and a combined validation set. The models are trained using a joined dictionary capped at 32,768 tokens, while evaluation is conducted separately on the test sets (each has 2000 samples) of each domain.

Multilingual OPUS-16. The OPUS-16 benchmark (Zhao et al., 2024) comprises 16 diverse languages. We prepare the training, validation, and test data following Zhao et al. (2024). The training and validation data are filtered to remove sentences exceeding 100 tokens. We train a joint Sentence-Piece unigram model with a base vocabulary of 32,000 tokens on the training set and subsequently apply it to encode all data splits.

A.2 Architecture

The encoder-decoder backbone has $L_{\text{enc}}=L_{\text{dec}}=8$ layers, $d_{\text{model}}=512$, and $h=8$ heads ($d_h=64$). We use a Pre-Norm setup with RMSNorm (Zhang and Sennrich, 2019) applied to token embeddings, layer inputs, and attention queries/keys (QK-norm). RoPE (Su et al., 2024) position embeddings are used. MoE blocks with SwiGLU

(Shazeer, 2020) non-linearities replace *every* feed-forward layer, each containing $E=4$ routed Matryoshka experts with elastic inner dimensions $\mathcal{D} = \{256, 512, 1024, 2048\}$ plus one Direction-Aware Shared Expert (DASE). Both the router and DASE are conditioned on a 64-dimensional direction embedding. Source and target embeddings are tied with the output projection, and final logits are soft-capped at 30.0.

A.3 Hyperparameters

The hyperparameters are reported in Table 10.

A.4 Model Selection and Evaluation

We compute validation BLEU every 1,000 steps and select the checkpoint with the highest validation BLEU *averaged over the two translation directions* as the final model for evaluation. For sub-network selection, we apply the same criterion independently to each $D_k \in \mathcal{D}$ on the same validation set. For evaluation on PhoMT and WMT-14, we use sacreBLEU⁴ v2.4.0 (signature nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0). For multi-domain, we use tokenized BLEU with multi-bleu.pl⁵ matching the conventions of prior work.

Significance Testing. We conducted paired approximate randomization (10,000 trials) alongside paired bootstrap resampling (1,000 resamples) to establish 95% confidence intervals. These tests confirm that the performance gains of the EDWARD-NMT sub-networks over the baselines are highly robust and statistically significant, consistently achieving $p < 0.001$ across our primary benchmarks.

A.5 Baselines

We compare against standard dense Transformers (BASE/BIG) (Vaswani et al., 2017; Ott et al., 2018), mBART (Liu et al., 2020), and MoE methods: the Switch Transformer (Fedus et al., 2022) with a top-1 token-based routing (ST-MoE) and GShard (Lepikhin et al., 2021) with a top-2 token-based routing (GS-MoE). For Transformers (BASE/BIG), ST-MoE and GS-MoE, we reproduce the results using Fairseq⁶. For mBART, we report the result

⁴<https://github.com/mjpost/sacrebleu>

⁵<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

⁶<https://github.com/facebookresearch/fairseq/tree/nllb>

³<https://github.com/pytorch/fairseq/tree/master/examples/translation>

Parameter	Value	Search Space
Optimiser	AdamW ($\beta_1=0.9, \beta_2=0.98, \epsilon=10^{-8}$)	—
Weight decay	0.01	—
Peak learning rate	10^{-2}	$\{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$
Initial learning rate	10^{-6}	—
Final learning rate	10^{-4}	—
Warmup steps	1,000	$\{0, 1000, 2000, 4000\}$
Schedule	cosine to η_{end}	—
Total steps	16k (multi-domain), 30k (OPUS-16), 40k (PhoMT), 60k (WMT-14)	—
Effective batch (sequences)	8,192	—
Micro-batch / accum. steps	2,048 / 4	—
Max sequence length	128 tokens	—
Gradient clipping	global norm 1.0	—
Label smoothing	0.1	—
Dropout (attn / FFN)	0.1 / 0.1	—
Routed experts E	4	—
Shared experts	1	—
Max expert dim M	2,048	—
Nested widths $\mathcal{D}$	$\{256, 512, 1024, 2048\}$	—
Direction embedding dim	64	$\{32, 64, 128\}$
Router jitter η	0.05	$\{0.01, 0.05, 0.1\}$
Threshold range $[\tau_{\min}, \tau_{\max}]$	$[0.05, 0.60]$	$\tau_{\min} \in \{0.05, 0.1, 0.2\}$
Target sparsity	$\rho_{\text{target}}=1.25$	$\{1.0, 1.25, 1.5\}$
Auxiliary weight	$\lambda_{\text{sp}}=0.01$	$\{0.01, 0.1, 1.0\}$
Dual-consistency weight	$\lambda_{\text{dc}}=1.0$	$\{0.01, 0.1, 1.0\}$
Decoding	Beam search, beam 4, length penalty 0.6	—

Table 10: Full hyperparameter set and search space. Values marked with “—” were fixed based on hardware constraints or standard architectural practices. The same router and elasticity settings are used across all benchmarks.

from Doan et al. (2021) on PhoMT, and we use the pretrained mBART many-to-many model⁷.

A.6 Software and Hardware Settings

All of our main experiments on EDWARD-NMT are implemented in JAX⁸ (0.9.2) and Flax⁹ (0.12.5). Training uses FSDP-style sharding with `jax.distributed` and gradient accumulation through `jax.lax.scan` so that the effective batch of 8192 sequences fits on a single TPU v5-64 pod. The dual-consistency objective is implemented by sampling two width masks per step inside the scanned step function, so the two forward passes share both the parameter slice and the data batch. All training and inference was performed on 64 chips TPU v5e, with total training times of approximately 6 hours for the multi-domain benchmark, 9 hours for OPUS-16, 12 hours for PhoMT, and 18 hours for WMT-14.

Training Cost and Efficiency. Profiling natively on a TPU v5e pod reveals that EDWARD-NMT requires ~ 1160 ms per training step, roughly $2.2 \times$

the time of standard MoE baselines (ST-MoE, GS-MoE; ~ 523 ms) and $3.3 \times$ the time of TRANSFORMER BIG (~ 348 ms). This overhead stems entirely from the Dual-Consistency (DC) objective evaluating two capacity views simultaneously. However, this joint optimization is highly efficient: training the S, M, L, and XL variants individually without DC would total 1267 ms per step, strictly exceeding our single-run cost. By extracting an entire family of sub-networks simultaneously, the amortized training time per model is just $2.01 \times$ that of TRANSFORMER BASE (144 ms), nearly half the cost of training a standard MoE ($3.63 \times$).

A.7 Step-by-Step Reproduction Guide

A self-contained reproduction of the primary empirical results reported in Section 5 and Appendix Section B requires the following steps. All artefacts are publicly available, and no proprietary code or weights are needed.

- 1. Prepare data.** Following Section A.1 for preparing the data and tokenizer.
- 2. Implement architecture.** Construct an encoder-decoder Transformer with the structural settings in Section A.2, replacing every

⁷<https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>

⁸<https://github.com/google/jax>

⁹<https://github.com/google/flax>

other feed-forward layer with the EDWARD-NMT MoE block of Section 3 (Equations 17 and the surrounding definitions of $E_i(x; D_k)$, $\psi(D_k)$, and $E_{\text{shr}}(x; D_k)$).

3. **Implement router.** Use the sigmoid router with jitter, learned threshold, and STE as defined in Section 3.3, with the target sparsity $\rho_{\text{target}}^{(t)}$ defined in Section A.3. Add the auxiliary losses $\mathcal{L}_{\text{sp}}$ with weight $\lambda_{\text{sp}}=0.01$.
4. **Dual-consistency loss.** At every optimization step, sample two widths D_A, D_B with replacement from $\mathcal{D}$, compute two forward passes that share parameters, and add DC Loss defined in Equation 13 to the cross-entropy loss with $\lambda_{\text{dc}}=1$.
5. **Train.** Pre-tokenized the data with TensorFlow Datasets¹⁰ with predefined max-length in Section A.3. Optimise with the schedule and batching of Table 10 for the number of steps indicated per benchmark.
6. **Extract sub-networks.** After training, obtain the four {S,M,L,XL} variants from the *same* weights by setting D_k to 256, 512, 1024, 2048 respectively at inference and applying the prefix mask $m^{(D_k)}$ together with the deterministic re-scaling $\psi(D_k)$. No fine-tuning, distillation, or post-hoc adaptation is performed.
7. **Decode and evaluate.** Decode with beam search (beam 4, length penalty 0.6) and evaluate with settings listed in Section A.4.

B Additional Results

B.1 Computational Results

To evaluate deployment viability, we benchmarked the extracted sub-networks across both edge and server regimes (Table 11). For edge deployment, we profiled models on a consumer-grade CPU (Apple M4, 4 threads, FP16) using compiled LiteRT checkpoints at a batch size of $B = 1$. As shown in Table 11, the most compact variant, EDWARD-NMT-S, activates only 57M parameters and requires just 6.28 GFLOPs. At $B = 1$, the system is inherently memory-bound; the sparse routing machinery introduces unavoidable memory-movement overhead, resulting in an on-device

throughput (104.9 TPS) that is highly competitive with, though slightly below, the dense TRANSFORMER BASE (117.7 TPS). However, it still drastically outperforms the TRANSFORMER BIG baseline (48.0 TPS) and delivers a $2.2\times$ speedup over our full XL variant (47.6 TPS).

In server-side deployment (profiled natively via JAX on a single NVIDIA H100 GPU), this routing overhead is efficiently amortized across larger batches, allowing the theoretical FLOP advantages to materialize dramatically. From $B = 4$ onwards, EDWARD-NMT-S strictly outperforms TRANSFORMER BASE, ultimately delivering a $3.2\times$ throughput speedup at $B = 128$ (70,372 vs. 21,755 TPS). Similarly, the full-capacity XL variant is $5.3\times$ faster than TRANSFORMER BIG at $B = 128$ while activating significantly fewer parameters. Ultimately, our train-once architecture effectively balances high representational capacity with the strict latency requirements of edge deployment, while concurrently offering massive scalability for high-throughput servers.

B.2 Multilingual Results

Baselines. We report our model on OPUS-16 in Table 12. For the MoE baselines, we report: Switch Transformers (Fedus et al., 2022) with a top-1 token-based routing (as ST-MoE); GShard (Lepikhin et al., 2021) a top-2 token-based routing (as GS-MoE); LS-MoE with fixed routing (Pires et al., 2023); HYBRID-MoE (Kudugunta et al., 2021) with a top-2 token routing in the encoder and a top-2 target language routing in the decoder; RESIDUAL-MoE (Elbayad et al., 2023; Rajbhandari et al., 2022b; Zhang et al., 2021) that augments each MoE layer with a shared feed-forward network through a binary gate function; LINGUAL-SMoE (Zhao et al., 2024) as MoE with linguistic-guided routing and dynamic expert allocation. We acquire these baselines result from Zhao et al. (2024). We also report THOR-MoE result from Liang et al. (2025).

Settings. For comparison, we report our EDWARD-NMT result with 2 settings: (1) same settings as described in Section 4.2; (2) we reduce the number of encoder and decoder layers to $L=L_{\text{enc}}=L_{\text{dec}}=6$ to match other baseline settings. Both settings use the same training hyperparameters in Section A.3. We use temperature-based data sampling strategy with $T = 1.5$. We report the model parameters from two settings in Table 13.

¹⁰<https://www.tensorflow.org/datasets>

Model	Params	Active	FLOPs (G)	Edge	Server Throughput (TPS)				
				TPS ($B = 1$)	$B = 1$	$B = 4$	$B = 16$	$B = 64$	$B = 128$
TRANSFORMER BASE	61M	61M	7.98	117.7	2,354	5,867	13,078	19,694	21,755
TRANSFORMER BIG	210M	210M	31.26	48.0	1,682	3,907	8,037	10,607	11,114
EDWARD-NMT-S	73M	57M	6.28	<u>104.9</u>	<u>1,763</u>	6,640	20,418	48,022	70,372
EDWARD-NMT-M	105M	72M	8.83	<u>88.7</u>	<u>1,688</u>	5,900	20,309	47,669	68,786
EDWARD-NMT-L	168M	102M	14.07	61.5	1,488	5,650	19,364	44,270	64,766
EDWARD-NMT-XL	294M	161M	25.20	47.6	1,410	5,218	17,499	41,856	59,199

Table 11: Inference efficiency across edge and server deployment regimes (64 output tokens). FLOPs (lower is better) are measured via native CPU execution. Edge throughput (Tokens Per Second, TPS, higher is better) is profiled using compiled LiteRT checkpoints on an Apple M4 CPU (4 threads, FP16) to simulate strict on-device constraints ($B = 1$). Server throughput is measured via JAX on a single NVIDIA H100 GPU across multiple batch sizes (B).

Models		En→XX				XX→En				Avg.
		high	medium	low	Avg1.	high	medium	low	Avg2.	
<i>Dense</i>	TRANSFORMER-BASE [†]	25.37	39.12	14.78	26.92	28.81	39.24	24.21	30.67	28.79
<i>MoE</i>	GS-MoE [†]	25.55	40.71	18.97	28.28	28.77	41.70	29.25	32.31	30.29
	ST-MoE [†]	26.55	43.04	20.23	29.68	29.94	44.00	30.94	33.89	31.79
	LS-MoE [†]	20.06	37.31	11.72	22.99	23.69	42.99	32.01	30.50	26.75
	HYBRID-MoE [†]	24.56	31.70	13.38	24.23	29.59	41.74	27.81	32.48	28.35
	RESIDUAL-MoE [†]	26.52	43.49	21.58	30.06	29.84	43.94	31.25	33.88	31.97
	LINGUAL-SMoE [†]	27.11	46.24	23.34	31.46	29.87	43.49	32.02	33.96	32.71
	THOR-MoE (Top-1) [‡]	27.21	46.37	23.88	31.17	29.68	44.01	32.58	33.99	32.57
	THOR-MoE (Top-2) [‡]	27.88	47.29	24.85	31.98	30.38	44.52	33.26	34.64	33.31
	THOR-MoE (Top- p) [‡]	<u>27.63</u>	46.82	24.15	31.55	29.97	44.18	32.96	34.26	32.91
	EDWARD-NMT-S	24.95	44.97	21.26	29.55	29.89	43.13	27.76	32.99	31.27
<i>Ours</i> ($L = 6$)	EDWARD-NMT-M	25.81	46.31	23.07	30.73	30.66	44.76	32.78	34.84	32.78
	EDWARD-NMT-L	26.31	47.53	24.67	31.64	31.34	45.83	34.45	35.82	33.73
	EDWARD-NMT-XL	26.76	48.22	25.68	32.27	31.77	46.49	34.79	36.30	34.28
<i>Ours</i> ($L = 8$)	EDWARD-NMT-S	25.78	46.34	23.52	30.81	30.81	44.46	33.20	34.93	32.87
	EDWARD-NMT-M	26.41	47.38	25.43	31.81	31.42	45.80	35.00	35.97	33.89
	EDWARD-NMT-L	27.11	<u>48.67</u>	<u>26.70</u>	<u>32.78</u>	<u>32.07</u>	<u>46.70</u>	<u>36.18</u>	<u>36.79</u>	<u>34.79</u>
	EDWARD-NMT-XL	27.26	49.27	28.15	33.30	32.42	47.43	36.65	37.27	35.29

Table 12: Averaged BLEU scores on multilingual OPUS-16 dataset. The **best** and second best scores in each column are highlighted in bold and underlined. [†]Reported from Zhao et al. (2024). [‡]Reported from Liang et al. (2025).

Noted that all the MoE baselines use 32 experts per MoE layers, MoE frequency is set to 2, with a total of $32 \times 6 = 192$ experts. Our second setting with 4 routed experts and 1 shared expert, result in $5 \times 12 = 60$ experts in total.

Variant	$L = 8$		$L = 6$	
	Params	Active	Params	Active
S	74M	57M	59M	47M
M	105M	72M	83M	59M
L	169M	102M	131M	83M
XL	295M	161M	226M	128M

Table 13: EDWARD-NMT variants with two settings of 8 layers ($L = 8$) and 6 layers ($L = 6$) on OPUS-16.

Results. As shown in Table 12, EDWARD-NMT establishes a new state-of-the-art on the multilingual OPUS-16 benchmark. When reducing the model depth to $L = 6$ to align with the scale of the baselines, the full EDWARD-NMT-XL variant achieves the highest overall average of 34.28 BLEU, comfortably outperforming THOR-MoE (33.31) and LINGUAL-SMoE (32.71). Our backbone achieves this superior quality with extreme expert efficiency, instantiating significantly fewer total experts (60 vs 192 in the baselines). Deploying our primary $L = 8$ architecture further widens the gap; EDWARD-NMT-XL dominates almost all individual resource brackets across both translation directions, peaking at an impressive average BLEU

of 35.29. Furthermore, our extracted sub-networks demonstrate exceptional parameter efficiency: even the highly compressed EDWARD-NMT-S variant at $L = 8$ (32.87 Avg) easily outperforms the dense TRANSFORMER-BASE baseline and remains highly competitive against much heavier sparse models.

B.3 Architecture Elasticity

Dim	P.	Act.	Sp.	En→Tgt	Tgt→En
PhoMT (En↔Vi)					
2048	287M	158M	1.44	43.79	39.13
1536	224M	128M	1.44	<u>43.62</u>	<u>38.98</u>
1024	161M	97M	1.45	43.61	38.93
768	130M	82M	1.46	43.48	38.58
512	98M	67M	1.47	43.35	38.50
256	67M	51M	1.51	42.91	37.88
128	51M	44M	1.56	39.26	33.98
WMT-14 (En↔De)					
2048	294M	175M	1.63	28.03	32.96
1536	231M	142M	1.63	<u>28.02</u>	<u>32.67</u>
1024	168M	109M	1.65	27.96	32.65
768	137M	92M	1.66	27.66	32.46
512	105M	76M	1.68	27.49	32.45
256	74M	59M	1.71	26.85	31.94
128	58M	51M	1.76	22.90	27.78
Multi-Domain (En↔De)					
2048	294M	161M	1.37	38.78	42.71
1536	231M	132M	1.37	<u>38.39</u>	<u>42.53</u>
1024	168M	102M	1.37	38.24	42.33
768	136M	87M	1.38	37.77	41.90
512	105M	72M	1.38	37.55	41.56
256	74M	57M	1.39	36.51	40.78
128	58M	50M	1.45	31.42	35.88

Table 14: Performance across dimensions on the PhoMT, WMT-14, and Multi-domain benchmarks. Column abbreviations denote total parameters (**P.**), average active parameters per token (**Act.**), and routing sparsity (**Sp.**) (active routed experts per token out of 4, excluding the shared expert). Multi-Domain scores are macro-averaged across five domains. The **best** and second best scores within each dataset are highlighted.

Table 14 provides the comprehensive empirical results supporting the structural elasticity analysis discussed in the main text. We report the exact translation quality, active parameter counts, and routing sparsity of EDWARD-NMT across all three benchmarks (PhoMT, WMT-14, and Multi-Domain) for both the explicitly trained dimensions ($\mathcal{D}$) and the unseen intermediate granularities. These detailed metrics empirically confirm the robust interpolation capabilities of our architecture within the optimized envelope (e.g., $D = 1536, 768$), as well as the expected performance breakdown when extrapolating below the minimum

trained capacity ($D = 128$).

B.4 Ablation on Variance Preservation

Setting	En↔De			
	XL	L	M	S
w/o $\psi(D_k)$	40.52	40.03	39.20	38.35
With ψ	40.74	40.26	39.39	38.72

Table 15: Ablation of the Variance Preservation factor $\psi(D_k)$ on the Multi-domain benchmark, averaged over both translation directions.

Table 15 presents the empirical results for the variance preservation mechanism discussed in the main text. By comparing the full configuration against a baseline lacking the deterministic rescaling factor $\psi(D_k)$, we observe a consistent degradation in translation quality across all nested capacities. Crucially, the performance gap widens as the sub-network size decreases, confirming that explicitly stabilizing the down-projected activation magnitudes is indispensable for maintaining the structural integrity of highly compressed dimensions (e.g., $D = 256$) during joint training.

B.5 Detailed Multi-domain Results

In Section 5.1, we reported the performance of EDWARD-NMT on the multi-domain benchmark averaged across both translation directions (De↔En) to provide a concise overview of model efficiency. For completeness, Table 16 provides the detailed, uni-directional tokenized-BLEU scores for all five domains (KOR, IT, MED, LAW, SUB) separately for the De→En and En→De directions. Consistent with the averaged results, our extracted sub-networks (L and XL) continue to demonstrate state-of-the-art performance against established baselines on almost all individual domains while requiring significantly fewer active parameters.

B.6 Additional WMT-14 Results

To provide a comprehensive evaluation of character-level translation quality, we additionally report the chrF2 metric on the WMT-14 benchmark (Table 17). The chrF2 results align perfectly with our primary BLEU and COMET evaluations. Specifically, the highly compressed EDWARD-NMT-S variant (59M active) substantially outperforms the comparably sized TRANSFORMER BASE (61M active) by +1.70 and +2.20 points in the En→De and De→En directions, respectively. At the upper bound, EDWARD-NMT-

Method	Params	Active	De→En						En→De					
			KOR	IT	MED	LAW	SUB	AVG	KOR	IT	MED	LAW	SUB	AVG
TRANSFORMER BASE	61M	61M	20.95	42.66	53.04	56.90	27.16	40.14	20.10	37.35	49.45	50.53	23.42	36.17
TRANSFORMER BIG	210M	210M	22.06	44.99	<u>55.37</u>	59.36	28.75	42.11	23.03	38.68	51.32	53.08	24.70	38.16
ST-MoE	360M	210M	22.96	43.90	<u>54.27</u>	58.79	28.30	41.64	20.49	38.10	49.75	52.59	24.06	37.00
GS-MoE	360M	260M	20.36	<u>44.79</u>	54.23	57.51	25.99	40.58	21.03	37.91	50.38	51.60	23.44	36.87
EDWARD-NMT-S	74M	57M	22.27	42.67	53.13	57.08	28.76	40.78	20.94	37.47	48.89	50.89	24.36	36.51
EDWARD-NMT-M	105M	72M	22.93	43.46	54.07	58.29	29.07	41.56	22.51	38.27	50.14	52.13	24.68	37.55
EDWARD-NMT-L	168M	102M	23.54	44.36	55.01	59.44	29.31	<u>42.33</u>	23.58	38.69	50.88	<u>53.09</u>	24.94	38.24
EDWARD-NMT-XL	294M	161M	23.65	44.69	55.59	60.21	29.43	42.71	24.46	39.10	<u>51.31</u>	53.89	25.16	38.78

Table 16: Detailed multi-domain results on the De↔En benchmark with tokenized-BLEU.

Method	Active	WMT-14 chrF2 (↑)	
		En→De	De→En
TRANSFORMER BASE	61M	55.82	56.23
TRANSFORMER BIG	210M	58.20	58.38
mBART [†]	610M	56.42	58.85
EDWARD-NMT-S	59M	57.52	58.43
EDWARD-NMT-M	76M	58.08	58.83
EDWARD-NMT-L	109M	<u>58.45</u>	<u>59.02</u>
EDWARD-NMT-XL	175M	58.59	59.29

Table 17: Character n -gram F-score (chrF2, higher is better) on the WMT-14 benchmark. Best and second-best scores are **bolded** and underlined, respectively. [†]mBART is a pretrained many-to-many model.

Method	Active	TER (↓)	
		En→Vi	Vi→En
TRANSFORMER BASE	65M	43.79	44.28
TRANSFORMER BIG	213M	43.04	44.06
mBART [†]	448M	42.57	<u>42.54</u>
EDWARD-NMT-S	51M	42.75	43.27
EDWARD-NMT-M	66M	42.43	42.88
EDWARD-NMT-L	97M	<u>42.21</u>	42.56
EDWARD-NMT-XL	158M	42.13	42.42

Table 18: Translation Edit Rate (TER, lower is better) on the PhoMT benchmark. [†]mBART is a pretrained many-to-many model.

XL achieves the highest overall chrF2 scores, robustly surpassing both TRANSFORMER BIG and the 610M-parameter mBART despite activating significantly fewer parameters.

B.7 Additional PhoMT Results

Post-editing Efficiency To evaluate the practical post-editing effort required for our generated translations, we report the Translation Edit Rate (TER, lower is better) on the PhoMT benchmark (Table 18). The results strictly corroborate our BLEU improvements and highlight the extreme parameter efficiency of our architecture. Despite

Setting	En↔Vi			
	XL	L	M	S
No loss	<u>41.41</u>	<u>41.19</u>	<u>40.82</u>	<u>40.13</u>
R-Drop	41.32	41.01	40.50	39.62
Asym. KL	41.31	41.13	40.65	40.08
DC Loss (ours)	41.46	41.27	40.93	40.40

Table 19: Ablation on the Dual-Consistency loss (DC Loss) on PhoMT, averaged over both directions.

Setting	En↔Vi			
	XL	L	M	S
MoE ($k=2$)	40.72	40.71	40.56	39.86
+ ME	41.31	41.15	40.69	40.04
+ ME + Sigmoid	<u>41.41</u>	41.19	40.82	40.13
+ ME + DC	41.38	<u>41.23</u>	<u>40.90</u>	40.42
+ ME + DC + Sigmoid	41.46	41.27	40.93	<u>40.40</u>

Table 20: Ablation on the contributions of each EDWARD-NMT component (PhoMT), averaged over both translation directions. ME, DC, Sigmoid stand for Matryoshka Experts, Dual-Consistency and Adaptive Sigmoid Routing respectively. *All settings use DASE.*

utilizing $4\times$ fewer active parameters, our most compressed sub-network, EDWARD-NMT-S (51M active), achieves a lower TER than TRANSFORMER BIG (213M active) in both translation directions. Furthermore, the full-capacity EDWARD-NMT-XL (158M active) achieves the lowest overall post-editing effort (42.13 En→Vi and 42.42 Vi→En), effectively outperforming the massively parameterized mBART model (448M) while operating with nearly $3\times$ fewer active parameters.

Additional Ablations on Consistency Strategies.

Table 19 provides additional ablation results for the consistency loss variants on the PhoMT benchmark (averaged over En↔Vi), complementing the multi-domain findings in the main text. The observed trends align perfectly with our primary results: applying the symmetric R-Drop objective degrades

performance across all scales. While asymmetric KL distillation (using the largest variant as a detached teacher) mitigates this severe degradation, it still falls slightly short of the baseline. Consistent with the multi-domain experiments, our proposed Dual-Consistency (**DC Loss**) provides the optimal regularization strategy, successfully outperforming the “No loss” baseline and achieving the highest scores across all model capacities (XL, L, M, S).

Additional Ablations on Contribution. Table 20 incrementally adds the EDWARD-NMT components on top of a standard Top-2 MoE baseline on PhoMT (noting that all settings use DASE). Adding Matryoshka Experts (ME) yields immediate improvements across all widths (e.g., +0.59 BLEU on XL and +0.18 on S over the Top-2 baseline). The gains are further amplified by the Dual-Consistency (DC) objective, which acts as a crucial regularizer to keep the smaller sub-networks competitive (+0.38 BLEU on S over ME alone). Finally, replacing the fixed Top- $k=2$ routing with our proposed Adaptive Sigmoid Router provides the best overall performance for the larger variants (XL, L, M).

B.8 Additional Hyperparameter Ablations

Direction	Target Sparsity (ρ_{target})		
	1.0	1.25	1.5
Avg.	40.40	40.66	40.73

Table 21: Ablation on target density (ρ_{target}) for the XL variant on the multi-domain benchmark.

Direction	Min Threshold (τ_{min})		
	0.01	0.05	0.10
Avg.	40.63	40.66	40.34

Table 22: Ablation on the min routing threshold (τ_{min}) for the XL variant on the multi-domain benchmark.

Direction	Router Jitter (η)		
	0.01	0.05	0.10
Avg.	38.98	40.66	40.62

Table 23: Ablation on the router jitter noise (η) for the XL variant on the multi-domain benchmark.

Other additional ablations on hyperparameters. We train only the largest variant EDWARD-NMT-XL ($D = 2048$), without DC Loss. We report

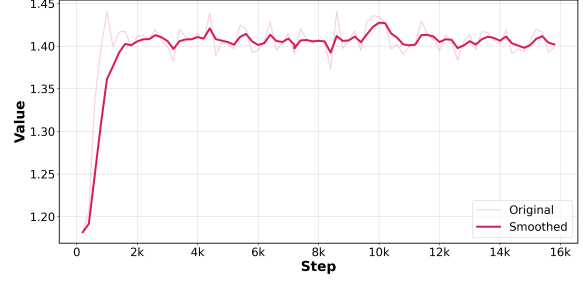


Figure 2: Average active experts per token during training.

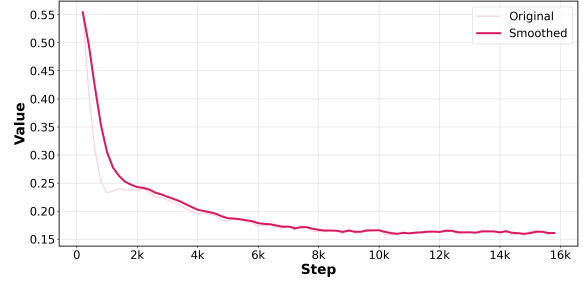


Figure 3: Learnable threshold τ over the course of training.

Setting	En $\leftrightarrow$ De			
	XL	L	M	S
$\lambda_{dc} = 0.1$	40.98	40.26	39.44	38.34
$\lambda_{dc} = 1.0$	40.75	40.28	39.56	38.65
$\lambda_{dc} = 5.0$	38.33	38.23	37.92	37.44

Table 24: Ablation on the Dual-Consistency loss weight (λ_{dc}) on the multi-domain benchmark, averaged over both translation directions.

Method	Target Density (ρ)		
	$\rho = 1.0$	$\rho = 1.25$	$\rho = 1.5$
EDWARD-NMT-S	1.18	1.39	1.69
EDWARD-NMT-M	1.17	1.38	1.67
EDWARD-NMT-L	1.16	1.37	1.66
EDWARD-NMT-XL	1.15	1.37	1.65

Table 25: Average active routed experts per token on the Multi-domain benchmark across different target density (ρ) settings. Values are averaged across both translation directions and all domains.

the evaluation result of the hyperparameters in Table 21, Table 22 and Table 23.

We report the active experts and the dynamic threshold of the model during training in Figure 2 and Figure 3. The number of active experts per layer in encoder and decoder of each variant is reported in Figure 4.

Table 24 shows the impact of the consistency

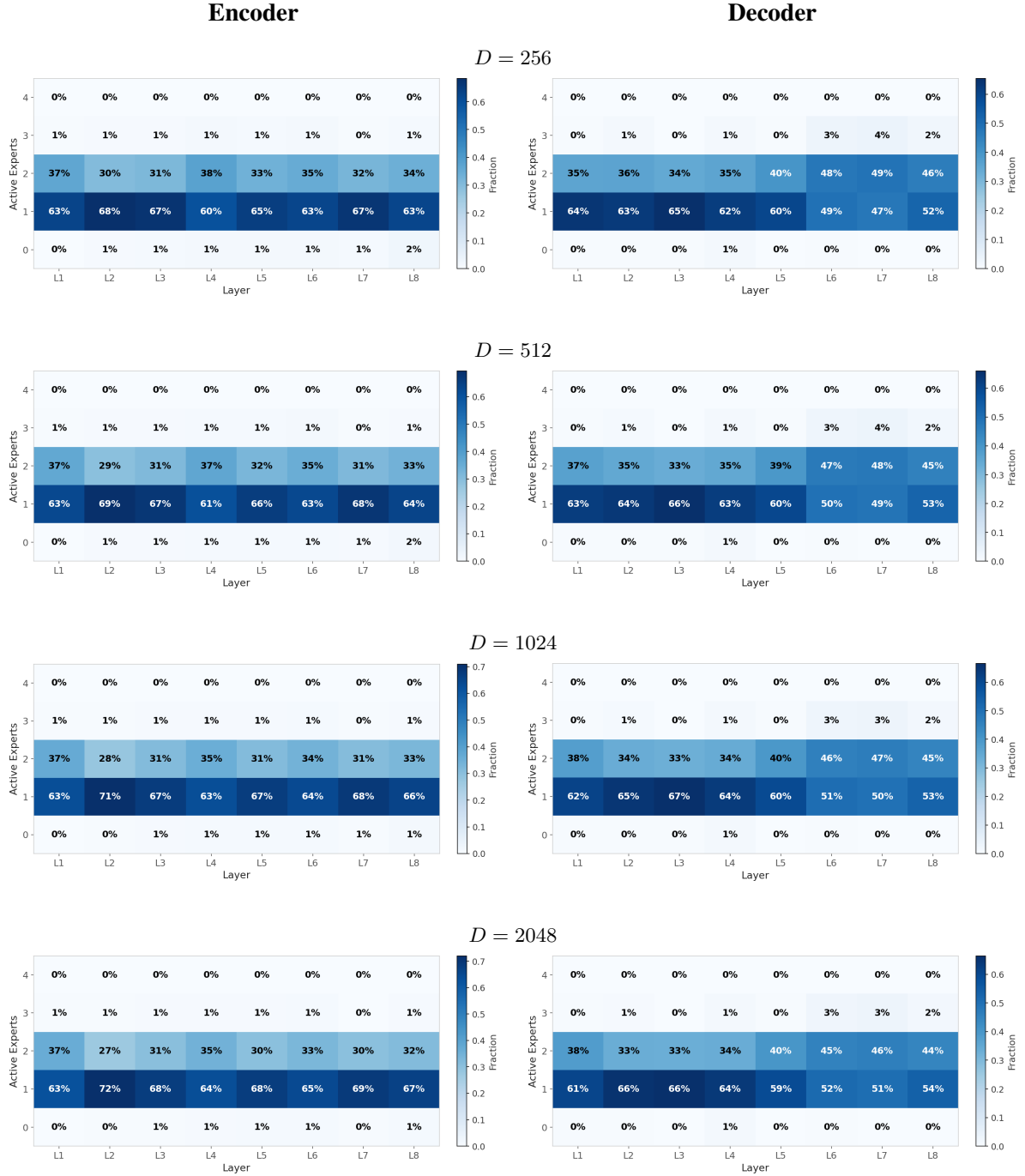


Figure 4: Percentage of number of active experts per layer across different elastic dimensions on the multi-domain benchmark.

weight λ_{dc} . A smaller weight ($\lambda_{dc} = 0.1$) favors the full-capacity XL model but yields sub-optimal smaller variants due to insufficient knowledge transfer. Conversely, an excessively large weight ($\lambda_{dc} = 5.0$) over-regularizes and degrades all scales. Our default $\lambda_{dc} = 1.0$ provides the optimal trade-off, maximizing the quality of the deployment-critical L, M, and S sub-networks.

adjusting the target density ρ , the empirical active experts scale predictably across all capacities (e.g., from ~ 1.15 experts at $\rho = 1.0$ up to ~ 1.69 at $\rho = 1.5$). This confirms that our auxiliary loss successfully grants fine-grained, predictable control over the inference compute budget without requiring any architectural modifications.

Table 25 validates our sparsity constraint. By