SPRI: SVD-Partitioned Residual Initialization for Data-Constrained MoE Upcycling

Anonymous ACL submission

Abstract

Mixture-of-Experts (MoE) models enable efficient scaling, but training them from scratch remains prohibitively expensive. MoE upcycling mitigates this cost by converting pretrained dense models into sparse MoE models. However, existing upcycling methods typically rely on large-scale continued training and often perform poorly under data-constrained supervised adaptation, due to either homogeneous experts or overly disruptive perturbations to pretrained parameters. In this setting, effective upcycling must leverage pretrained weight structure while introducing sufficient diversity among routed experts. To this end, we propose SVD-Partitioned Residual Initialization (SPRI), which distributes SVD-partitioned residuals derived from pretrained feed-forward network (FFN) weights across routed experts, introducing controlled expert diversity grounded in pretrained spectral structure. We further introduce a two-stage training strategy to improve adaptation stability. We evaluate SPRI on multilingual speech-to-text translation, where limited supervised data challenges MoE upcycling and multiple target languages provide natural routing heterogeneity. On CoVoST2 across 15 En-to-XX directions, SPRI improves average BLEU and COMET over fully fine-tuned dense models by 2.58 and 3.32 points, respectively, and outperforms the prior best MoE upcycling baseline by 2.01 BLEU and 3.04 COMET points.

1 Introduction

Large language models (LLMs) with Mixture-of-Experts (MoE) architectures have emerged as a practical paradigm for scaling model capacity without proportionally increasing per-token computation, by activating only a sparse subset of parameters for each input (Jiang et al., 2024; Dai et al., 2024; Muennighoff et al., 2025). However, training MoE models from scratch remains prohibitively expensive. MoE upcycling offers a more efficient alternative by converting a pretrained dense model into a sparse MoE model, thereby reusing dense pretrained knowledge as a strong initialization for expert-based scaling (Komatsuzaki et al., 2022; He et al., 2024).

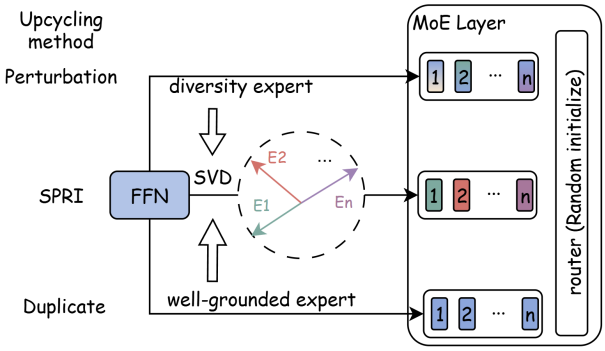


Figure 1: Overview of SVD-Partitioned Residual Initialization (SPRI). Compared with conventional upcycling methods that duplicate FFN weights or add unstructured perturbations, SPRI assigns structured spectral residuals to routed experts, yielding controlled expert diversity grounded in pretrained weight structure.

Most existing MoE upcycling methods initialize each MoE layer by duplicating the feed-forward network (FFN) of the corresponding dense Transformer block across multiple routed experts (Komatsuzaki et al., 2022; He et al., 2024). Recent open-source MoE models further introduce random perturbations or partial re-initialization to break expert symmetry and encourage specialization (Yang et al., 2024; Muennighoff et al., 2025; Nakamura et al., 2025). However, these strategies are typically paired with large-scale continued pretraining, often involving hundreds of billions or even trillions of tokens. This leaves an important question underexplored: *can MoE upcycling still provide effective capacity expansion when only limited adaptation data is available?*

In this limited-data regime, effective MoE upcycling requires more than parameter reuse: it must leverage the internal structure of pretrained

FFN weights to create diverse yet well-grounded routed experts. Directly copying the dense FFN provides a conservative initialization from the pretrained model, but produces nearly identical experts and limits the benefit of sparse expert modeling (Nakamura et al., 2025). As we show empirically, perturbation-based strategies can introduce expert diversity, but their unstructured modifications often make optimization less reliable in the absence of large-scale continued pretraining. Thus, the key challenge is to construct expert heterogeneity from the pretrained model itself, rather than relying on unstructured perturbations.

To address this challenge, we propose **SVD-Partitioned Residual Initialization** (SPRI), an MoE upcycling method based on singular value decomposition (SVD). SPRI decomposes pretrained FFN weights and assigns structured spectral residuals from distinct subspaces to different routed experts. This produces bounded and interpretable expert variation while keeping the dense FFN as a functional anchor. We further introduce a two-stage training strategy that first adapts only the residual components and then finetunes all routed expert parameters, making MoE upcycling more stable under limited data.

We evaluate SPRI on multilingual speech-to-text translation (S2TT), a practical setting where additional expert capacity is desirable but large-scale continued pretraining is often infeasible. Multilingual translation benefits from expert specialization across languages and language pairs (Gu et al., 2018; Kudugunta et al., 2021; Zhao et al., 2024), while supervised speech-translation data remains substantially smaller than typical continued-pretraining corpora (**?**Wang et al., 2021). On 15 English-to-X directions in CoVoST2, SPRI improves over fully fine-tuned dense models by 2.58 BLEU and 3.32 COMET points on average, and further outperforms the prior best MoE upcycling baseline by 2.01 BLEU and 3.04 COMET points.

2 Method

2.1 Mixture of Experts

Sparse MoE layers routing each token to a small subset of experts. Formally, given an input $\mathbf{x} \in \mathbb{R}^d$, each expert $\mathbf{E}_i$ in an MoE layer is implemented as a gated MLP:

$$\mathbf{E}_i(\mathbf{x}) = \mathbf{W}_i^{\text{Down}} \left(\sigma \left(\mathbf{W}_i^{\text{Gate}} \mathbf{x} \right) \odot \mathbf{W}_i^{\text{Up}} \mathbf{x} \right), \tag{1}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ is the activation function.

At layer l, the router assigns token to the routed experts and activates the Top-k experts in E routed experts, where $k \ll E$:

$$\begin{aligned} \mathbf{g}^l(\mathbf{x}) &= \text{Softmax} \left(\left\{ (\mathbf{W}^l_{\text{router}} \mathbf{x})_i \right\}_{i \in \mathcal{S}^l_k(\mathbf{x})} \right), \\ \mathcal{S}^l_k(\mathbf{x}) &= \text{TopK} \left(\mathbf{W}^l_{\text{router}} \mathbf{x}, k \right). \end{aligned} \tag{2}$$

Following the recent MoE architectures, which introduce shared experts that are always activated (Dai et al., 2024; DeepSeek-AI, 2024), the MoE layer is then computed as:

$$\text{MoE}^l(\mathbf{x}) = \sum_{i \in \mathcal{S}^l_k(\mathbf{x})} g^l_i(\mathbf{x}) \mathbf{E}^l_i(\mathbf{x}) + \mathbf{E}^l_{\text{shared}}(\mathbf{x}). \tag{3}$$

2.2 Preliminary Analysis of MoE Upcycling Strategies

We first review conventional MoE upcycling strategies and formulate them under a unified notation.

Naive Upcycling. Naive upcycling (Komatsuzaki et al., 2022; He et al., 2024) converts a pretrained dense model into a sparse MoE model by initializing all routed experts with the original dense MLP module and randomly initializing the router. For each expert $i \in \{1, \ldots, E\}$ in naive upcycling, the initialization is:

$$\mathbf{W}_i^{(\cdot)} = \mathbf{W}_{\text{MLP}}^{(\cdot)}, \tag{4}$$

where $(\cdot) \in \{\text{Gate}, \text{Up}, \text{Down}\}$.

Noise Upcycling. Noise upcycling (Chu et al., 2024b)[1] breaks the symmetry among duplicated experts by adding random perturbations to a subset of copied weights. For each expert i, the initialized weight matrix is:

$$\mathbf{W}_i^{(\cdot)} = \mathbf{W}_{\text{MLP}}^{(\cdot)} + (\mathbf{1} - \mathbf{M}_i^{(\cdot)}) \odot \boldsymbol{\epsilon}_i^{(\cdot)}, \tag{5}$$

where $\mathbf{M}_i^{(\cdot)}$ is a binary mask indicating the copied parameters, and each entry of $\boldsymbol{\epsilon}_i^{(\cdot)}$ is independently sampled from $\mathcal{N}(0, \sigma^2)$.

[1] Since the Qwen2 technical report does not fully specify the perturbation procedure, we follow the descriptions in (Muennighoff et al., 2025; Nakamura et al., 2025).

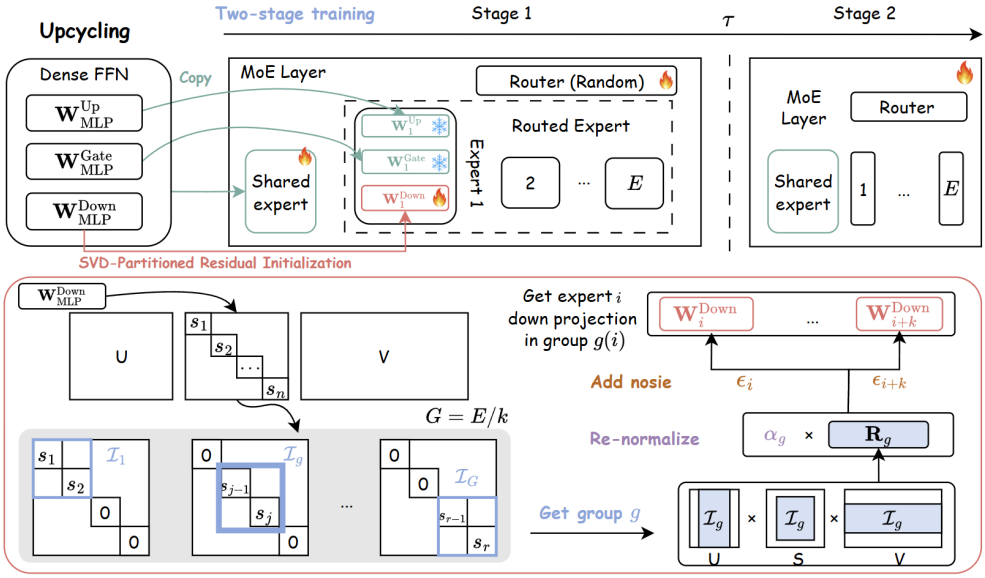


Figure 2: Pipeline of SVD-Partitioned Residual Upcycling. The shared expert copies the dense MLP to preserve pretrained behavior, while routed experts receive norm-controlled residual down projections constructed from disjoint SVD partitions.

Drop Upcycling. Drop upcycling (Nakamura et al., 2025) promotes expert diversity by partially reinitializing the MLP parameters of duplicated experts. It randomly selects a subset of indices along the intermediate dimension for each copied expert and reinitializes the corresponding rows or columns of the MLP weight matrices, while leaving the remaining parameters unchanged:

$$\mathbf{W}_i^{(\cdot)} = \mathbf{M}_i^{(\cdot)} \odot \mathbf{W}_{\text{MLP}}^{(\cdot)} + \bar{\mathbf{M}}_i^{(\cdot)} \odot \mathbf{R}_i^{(\cdot)}, \quad (6)$$

where $\mathbf{M}_i^{(\cdot)}$ denotes a structured binary mask induced by the selected intermediate dimensions, $\bar{\mathbf{M}}_i^{(\cdot)} = \mathbf{1} - \mathbf{M}_i^{(\cdot)}$ is its complementary mask, and $\mathbf{R}_i^{(\cdot)}$ denotes newly initialized parameters.

Branch-Train-MiX (BTX). BTX (Sukhbaatar et al., 2024) first trains multiple dense models on different domain-specific datasets and then merges their MLP module as routed experts, while averaging the remaining shared parameters:

$$\mathbf{W}_i^{(\cdot)} = \mathbf{W}_{\text{MLP},i}^{(\cdot)}, \quad (7)$$

where $\mathbf{W}_{\text{MLP},i}^{(\cdot)}$ denotes the corresponding MLP weight from the i-th branch-trained dense model.

Preservation and Diversity Analysis. We analyze how initialization impacts knowledge preservation and expert diversity in MoE upcycling.[2]

[2]BTX is excluded from this analysis because it requires additional branch training and is not a direct single-model upcycling method.

Method	KL ↓	Diversity ↑	BLEU ↑
Dense FT	0	-	19.66
Naive Upcycling	1.64	0.01%	17.19
Noise Upcycling	1.64	2%	18.36
Drop Upcycling	4.93	99.99%	18.85
SPRI (ours)	0.12	88.02%	22.24

Table 1: Preservation and diversity comparison across MoE upcycling methods. Dense FT denotes the fully fine-tuned dense model. KL measures the output-distribution deviation between each upcycled model and the original dense model, diversity measures the discrepancy among routed experts, and BLEU denotes the average BLEU score across 15 En→XX translation directions. Statistical details in Appendix A.1.

As shown in Table 1, MoE architectures with higher diversity generally achieve higher BLEU scores, suggesting that expert diversity is beneficial for MoE architectures in our experimental setting. However, diversity alone is insufficient, all prior MoE upcycling baselines underperform the fully fine-tuned dense model with larger output KL divergence. This indicates that, when large-scale pretraining data is unavailable, effective MoE upcycling should introduce expert diversity while preserving the original dense model behavior as much as possible.

2.3 SVD-Partitioned Residual Upcycling

Motivated by the above analysis, we propose SPRI, which follows a dense-anchored residual design.

The shared expert preserves the pretrained dense MLP module, while routed experts introduce small, structured residual corrections. This design separates knowledge preservation from expert diversification, allowing the initialized MoE to remain close to the dense model while providing diversity among routed experts.

Instead of copying or perturbing the full MLP parameters for each routed expert, SPRI copies the gate and up projections to preserve the dense nonlinear feature transformation, and introduces expert-specific differences only through small SVD-partitioned residuals in the down projection. By controlling the residual magnitude, routed experts start as weak but structured corrections.

Dense-anchored shared expert. In specific MoE layer, the shared expert provides the anchor by fully copying the pretrained dense MLP module:

$$\mathbf{W}^{(\cdot)}_{\text{shared}} = \mathbf{W}^{(\cdot)}_{\text{MLP}}, \tag{8}$$

where $(\cdot) \in \{\text{Gate}, \text{Up}, \text{Down}\}$.

SVD-partitioned routed residuals. In each routed expert, the gate and up projections are initialized as:

$$\begin{aligned} \mathbf{W}^{(\text{Gate})}_i &= \mathbf{W}^{(\text{Gate})}_{\text{MLP}}, \\ \mathbf{W}^{(\text{Up})}_i &= \mathbf{W}^{(\text{Up})}_{\text{MLP}}, \quad i = 1, \ldots, E. \end{aligned} \tag{9}$$

The down projection is then used to introduce structured diversity among routed experts. Specifically, let the dense down-projection matrix be decomposed by singular value decomposition:

$$\mathbf{W}^{(\text{Down})}_{\text{MLP}} = \mathbf{U}\,\text{diag}(\mathbf{s})\mathbf{V}^\top, \tag{10}$$

where $\mathbf{s} = (s_1, \ldots, s_r)$ contains the singular values in descending order. We use the routing Top-k as the expert group size and set the number of residual groups to $G = E/k$. For expert i, we denote its group assignment by $g(i) = \lceil i/k \rceil$, where $i \in \{1, \ldots, E\}$ and $g(i) \in \{1, \ldots, G\}$.

Each group is associated with one spectral residual matrix. Specifically, we split the ordered singular-component indices $\{1, \ldots, r\}$ into G contiguous blocks $\{\mathcal{I}_g\}_{g=1}^{G}$. For group $g \in \{1, \ldots, G\}$, the corresponding residual matrix is reconstructed as:

$$\mathbf{R}^{(\text{Down})}_g = \mathbf{U}_{[:,\mathcal{I}_g]}\,\text{diag}\big((s_j)_{j\in\mathcal{I}_g}\big)\mathbf{V}^\top_{[:,\mathcal{I}_g]}. \tag{11}$$

The down projection of routed expert $i \in \{1, \ldots, E\}$ is initialized from the residual matrix of its group:

$$\mathbf{W}^{(\text{Down})}_i = \alpha_{g(i)}\mathbf{R}^{(\text{Down})}_{g(i)} + \boldsymbol{\epsilon}_i, \tag{12}$$

where $\boldsymbol{\epsilon}_i$ is a small expert-specific perturbation that breaks symmetry among experts in the same group.

While groups contain larger singular values tend to yield residual matrices $\mathbf{R}^{(\text{Down})}_{g(i)}$ with larger Frobenius norms. We introduce a group-wise scaling coefficient to re-normalize each residual matrices:

$$\alpha_g = \frac{\rho \left\| \mathbf{W}^{(\text{Down})}_{\text{MLP}} \right\|_F}{\left\| \mathbf{R}^{(\text{Down})}_g \right\|_F + \delta}, \tag{13}$$

where ρ controls the relative magnitude of each routed residual and δ is a small constant for numerical stability. A small ρ makes the initial MoE rely primarily on the shared expert as the preservation anchor.

Residual MoE interpretation. As a result, the initialized MoE layer can be viewed as the pretrained dense FFN plus a small routed residual term:

$$\begin{aligned} \text{SPRI}^l(\mathbf{x}) &= \mathbf{E}^l_{\text{shared}}(\mathbf{x}) + \sum_{i\in\mathcal{S}^l_k(\mathbf{x})} g^l_i(\mathbf{x})\Delta\mathbf{E}^l_i(\mathbf{x}) \\ \Delta\mathbf{E}^l_i(\mathbf{x}) &= \mathbf{W}^{(\text{Down})}_i \left(\sigma(\mathbf{W}^{(\text{Gate})}_{\text{FFN}}\mathbf{x}) \odot \mathbf{W}^{(\text{Up})}_{\text{FFN}}\mathbf{x}\right) \end{aligned} \tag{14}$$

2.4 Training Strategy

Two-stage training After initialization, we apply a two-stage update schedule to routed experts, to exploit the expert diversity.

During the first τT training steps, only $\{\mathbf{W}^{(\text{Down})}_i\}_{i=1}^{E}$ and the router are updated, while the copied gate and up projections are frozen. Afterward, $\{\mathbf{W}^{(\text{Gate})}_i, \mathbf{W}^{(\text{Up})}_i, \mathbf{W}^{(\text{Down})}_i\}_{i=1}^{E}$ are optimized jointly with the router. Here, T is the total number of training steps and τ is the freezing ratio.

Training objective We train all MoE models with the standard MoE training objective, consisting of the task cross-entropy loss and the auxiliary routing losses commonly used in sparse MoE models (Fedus et al., 2022; Zoph et al., 2022). The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{lb}}\mathcal{L}_{\text{lb}} + \lambda_z\mathcal{L}_z, \tag{15}$$

where $\mathcal{L}_{\text{CE}}$ is the task cross-entropy loss, $\mathcal{L}_{\text{lb}}$ is balance loss which encourages balanced expert utilization, and $\mathcal{L}_z$ is z-loss which regularizes router logits for stable training. The coefficients λ_{lb} and λ_z control the strengths of the corresponding auxiliary losses.

3 Experimental Setups

3.1 Datasets and Evaluation Metrics

We train our method on the S2TT task using the CoVoST 2 (Wang et al., 2020) and Europarl-ST (Iranzo-Sánchez et al., 2020), both large-scale multilingual speech translation datasets. We focus on the English-to-X setting to prevent the capability of the audio encoder from interfering with our experiments. We evaluate on all 15 target-language directions in the CoVoST 2 dataset. After preprocessing, the training split contains about 6.9k hours of speech.

We evaluate our method on CoVoST 2. We report translation quality using sacreBLEU (Post, 2018) and COMET (Rei et al., 2020). We follow the evaluation setting of Qwen2-Audio (Chu et al., 2024a) to compute all BLEU scores. For COMET, we use the `Unbabel/wmt22-comet-da`[3]. Additional details on preprocessing, data loading, and evaluation settings are provided in Appendix A.2.

3.2 Model Architecture

We conduct our main experiments using Qwen3-ASR-0.6B (Xian Shi, 2026), which consists of an acoustic encoder, a lightweight adapter that aligns acoustic features with the textual embedding space, and a decoder-only language model. Its moderate scale makes large-scale continued pretraining infeasible, which matches the limited-data scenario considered in this work. To construct sparse MoE variants, we convert the dense FFN modules in the language decoder into MoE layers at a regular interval of every four decoder layers. Each MoE layer contains 8 routed experts and 1 shared expert, and activates the Top-2 routed experts per token during both training and inference. For all MoE upcycling methods, we use the same shared-expert configuration, where the shared expert is initialized by copying the corresponding dense FFN. Different upcycling methods mainly differ in how they initialize the routed experts.

[3]https://huggingface.co/Unbabel/wmt22-comet-da

3.3 Training

During training, we keep all audio inputs resampled to 16 kHz and set the maximum sequence length to 768. We train each model for 1 epoch with a global batch size of 1024. Optimization is performed with AdamW using a peak learning rate of 2×10^{-5}, a linear learning rate schedule, and a warmup ratio of 0.02 with bf16 precision.

For MoE models, we use the same router regularization losses and coefficients, which $\lambda_{\text{lb}} = 10^{-2}$ and $\lambda_z = 10^{-3}$ across all upcycling methods.

For SPRI, we apply the two-stage training strategy described in Section 2.4, with $\tau = 0.1$, and we select $\rho = 10^{-3}$, $\delta = 10^{-12}$ for all MoE models. All experiments are conducted using 8 NVIDIA H800 GPUs with 80GB memory each.

4 Results

4.1 Main Results

Table 2 reports the main results on CoVoST2 English-to-X S2TT dataset. Conventional MoE upcycling baselines fail to consistently benefit from the increased parameter count in the data-constrained setting. Although these methods expand the model from 0.7B to 1.3B parameters, Naive, Drop, Noise, and BTX Upcycling all underperform Dense FT on average. This suggests that effective upcycling requires more than directly reusing or perturbing dense FFN parameters.

Among conventional upcycling methods, Drop and Noise improve over Naive Upcycling, indicating that expert diversity is more useful than homogeneous expert duplication. However, their performance remains below Dense FT, showing that unstructured perturbations are insufficient for reliable expert specialization. BTX also fails to close this gap, suggesting that decoupling expert construction from router adaptation is less effective than jointly optimizing routed experts and the router.

SPRI consistently outperforms all baselines. With two-stage training, it achieves 78.42 COMET and 22.24 BLEU, improving over Dense FT by +3.32 COMET and +2.58 BLEU and over the best conventional MoE upcycling baseline by +4.34 COMET and +3.39 BLEU. The single-stage variant also outperforms all baselines, indicating that the main gain comes from SVD-partitioned residual initialization, while two-stage training further improves performance. These results demonstrate that SPRI better balances pretrained knowledge

Model	Param	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Dense FT	0.7B	**77.28**	**12.46**	75.43	8.79	**86.01**	**30.36**	68.82	14.69	66.12	10.27	68.94	11.17	65.38	10.67	**80.75**	**28.52**
Lora	0.4B	74.89	11.02	**79.20**	**10.95**	84.73	28.93	**68.86**	**14.83**	**69.31**	**11.77**	**71.05**	**12.53**	**67.09**	**10.68**	78.25	24.92
Naive Upcycling	1.3B	73.94	10.36	75.18	8.33	84.41	28.35	65.51	12.43	64.18	8.85	68.01	10.55	62.47	8.56	77.99	24.60
Drop Upcycling	1.3B	75.94	11.56	74.60	8.46	85.34	29.14	67.63	14.26	65.28	9.81	67.35	10.41	64.29	10.17	79.45	27.23
Noise Upcycling	1.3B	75.37	10.98	73.87	8.11	85.12	28.77	66.91	13.63	64.51	9.53	66.76	10.18	63.38	9.92	78.89	26.58
BTX	1.3B	74.58	10.88	72.93	7.75	84.80	28.35	66.09	13.60	64.33	9.64	66.56	10.09	63.08	10.28	78.47	26.39
SPRI	1.3B	78.51	13.58	78.93	11.02	86.24	30.69	72.58	17.66	71.05	12.98	72.26	13.24	70.71	13.62	80.98	29.04
SPRI 2 stage	1.3B	**78.85**	**13.83**	**79.80**	**11.52**	**86.28**	**30.80**	**74.25**	**19.13**	**73.45**	**14.20**	**73.08**	**13.75**	**71.90**	**14.14**	**81.03**	**29.16**

Model	Param	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Dense FT	0.7B	**84.40**	**42.42**	67.62	4.37	**76.58**	**26.71**	67.82	16.40	**75.36**	**26.26**	**80.07**	**18.41**	**85.92**	**33.42**	75.10	**19.66**
Lora	0.4B	82.82	38.76	**74.24**	**7.11**	73.38	23.05	**70.73**	**18.42**	74.42	25.49	79.06	17.32	82.73	27.97	**75.38**	18.92
Naive Upcycling	1.3B	82.44	37.81	68.01	4.48	72.58	22.46	64.93	13.48	72.50	22.80	78.89	17.07	82.53	27.74	72.91	17.19
Drop Upcycling	1.3B	83.84	40.88	67.54	4.52	74.32	25.14	67.47	16.20	74.51	25.61	78.82	17.25	84.87	32.08	74.08	18.85
Noise Upcycling	1.3B	83.58	40.40	66.64	4.22	73.52	24.59	66.82	15.71	73.84	24.73	78.50	16.64	84.55	31.37	73.48	18.36
BTX	1.3B	83.17	39.69	66.70	4.31	73.22	24.46	66.78	15.63	73.55	24.70	77.88	16.53	84.13	31.38	73.09	18.24
SPRI	1.3B	84.56	42.85	73.77	**7.11**	77.72	28.15	73.36	22.22	77.10	29.30	80.41	18.80	86.12	33.89	77.62	21.61
SPRI 2 stage	1.3B	**84.57**	**43.04**	**75.57**	**7.94**	**77.98**	**28.52**	**75.05**	**24.06**	**77.81**	**30.42**	**80.46**	**19.03**	**86.26**	**34.14**	**78.42**	**22.24**

Table 2: Results on CoVoST2 English-to-X speech-to-text translation. We report COMET and sacreBLEU for each target language and their average over 15 directions. Dense FT denotes the fully fine-tuned dense model.

preservation and expert diversification than homogeneous copying or unstructured perturbations.

SPRI improves across all 15 translation directions, with smaller gains on high-resource languages such as Zh, De, and Ja. We conjecture that the pretrained dense model has already learned strong representations for these languages from higher-resource pretraining data, so the CoVoST2 fine-tuning signal provides relatively limited additional benefit. The degradation of conventional upcycling methods on these directions is consistent with this view: when the dense initialization is already strong, disrupting pretrained knowledge can be more harmful than the capacity gain from MoE upcycling.

To further verify the effectiveness of SPRI, we conduct a series of controlled ablation studies. Rather than tuning each component for the best possible score, these experiments aim to isolate the contribution of individual design choices. Unless otherwise specified, we vary one factor at a time while keeping the remaining settings fixed.

4.2 Ablation on SVD-Partitioned Routed Residuals

We further verify whether the gains of SPRI come from the SVD-partitioned routed residuals themselves. Table 3 reports two ablation experiments. In the first part, we only replace the routed down-projection initialization.

Copy, noise, and drop initialization achieve 19.64, 19.59, and 18.94 BLEU, respectively, while SVD-partitioned residuals achieve 22.24 BLEU and 78.42 COMET. This shows that the improvement is not simply due to applying upcycling only to the down projection or using the same training strategy. Instead, constructing routed experts from structured spectral directions is important for inducing useful expert diversity grounded in the pretrained FFN weights.

Variant	Avg. COMET	Avg. BLEU
Down-projection initialization		
Copy	75.16	19.64
Noise	75.11	19.59
Drop	74.23	18.94
SVD-Partitioned	**78.42**	**22.24**
Shared expert		
w/ shared expert	**78.42**	22.24
w/o shared expert	78.28	22.24

Table 3: Ablation on SVD-partitioned routed residuals. The first block keeps the SPRI framework fixed and varies only the routed down-projection initialization. The second block removes the shared expert to test whether SPRI relies on the dense preservation anchor. All direction detailed and experimental settings are provided in Appendix A.3.

The second part examines whether SPRI depends on the shared expert. Removing the shared expert gives the same BLEU as the default setting and only slightly decreases COMET from 78.42 to 78.28. This indicates that the main gain comes from SVD-partitioned routed residuals rather than from the shared expert alone. The shared expert can serve as a preservation anchor, but it is not necessary for SPRI to remain effective.

The second part examines whether SPRI depends on a permanently activated shared expert. Removing this permanent activation preserves the same BLEU and only slightly decreases COMET from 78.42 to 78.28. This suggests that SPRI's main gain comes from SVD-partitioned routed residuals, rather than from always activating a

dense-copied expert.

4.3 Parameter-Matched Scaling via MoE Upcycling

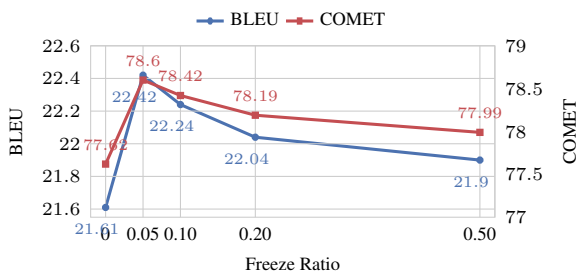
We further compare SPRI-based MoE upcycling with dense models at a similar total parameter scale. As shown in Table 4, SPRI-3B achieves 80.57 COMET and 24.43 BLEU, substantially outperforming both `Qwen2.5-Omni-3B` and its fine-tuned variant. Compared with `Qwen2.5-Omni-3B FT`, SPRI-3B improves the average score by +3.98 COMET and +4.48 BLEU.

These results suggest that, under data-constrained S2TT adaptation, expanding a smaller pretrained model through MoE upcycling can be more effective than directly fine-tuning a parameter-matched dense model. Since the upcycled MoE activates only a subset of experts per token, it also offers a more computation-efficient scaling path than dense models with a similar total parameter count.

Model	Avg. COMET	Avg. BLEU
`Qwen2.5-Omni-3B`	68.07	13.86
`Qwen2.5-Omni-3B FT`	76.59	19.95
SPRI-3B	**80.57**	**24.43**

Table 4: Parameter-matched comparison between SPRI-based MoE upcycling and dense models. At a comparable total parameter scale, SPRI-3B outperforms both `Qwen2.5-Omni-3B` and its fine-tuned variant on average COMET and BLEU. Full per-language results are provided in Appendix A.4.

4.4 Sensitivity to the Unfreezing Ratio



(a) Sensitivity to unfreezing ratio

Figure 3: Sensitivity to the freezing ratio τ in two-stage training. Full results provided in Appendix A.5.

We study the effect of the freezing ratio τ in two-stage training, where τ denotes the fraction of training steps during which the routed-expert gate and up projections remain frozen. As shown in Figure 3, a small nonzero freezing ratio improves over direct joint training: $\tau = 0.05$ achieves the best average performance, increasing COMET from 77.62 to 78.60 and BLEU from 21.61 to 22.42. This suggests that two-stage training benefits from a short frozen stage, which stabilizes early MoE upcycling, while over-freezing delays expert specialization.

4.5 Ablation on SVD Residual Placement

We study where to inject the SVD-partitioned residuals within the FFN module. Table 5 compares applying the residual construction to the up, down, and gate projections, while keeping all other settings unchanged. Applying the residual to the up projection gives the weakest result, suggesting that modifying the expansion projection is less effective for stable expert diversification. In contrast, both down and gate placement achieve stronger performance, showing that the location of spectral residuals has a clear impact on data-constrained MoE upcycling.

Although the gate variant obtains the highest score in this ablation, we keep down-projection placement as our default design.[4] We choose down placement because it preserves the copied gate/up projections as the dense nonlinear path and introduces routed residual corrections at the FFN output, matching our residual MoE formulation.

Placement	Avg. COMET	Avg. BLEU
Up	77.67	21.73
Down	78.42	22.24
Gate	**78.74**	**22.51**

Table 5: Ablation on SVD residual placement. Scores are averaged across all translation directions. Gate and down placement are both effective, while applying SVD residuals to the up projection performs worse. Full per-language results are provided in Appendix A.6.

4.6 Sensitivity to MoE Architecture Configurations

We study how MoE architecture choices affect SPRI-based upcycling in Table 6, including MoE layer placement, expert/activation configuration, and routing Top-k. The results show that effective upcycling depends not only on the amount of added expert capacity, but also on where it is inserted and how much of it is activated.

[4] This ablation is conducted under $\tau = 0.1$, while the main experiments use $\tau = 0.05$.

Upcycling more FFN layers consistently improves performance. In particular, the every-layer setting outperforms the 32-expert/Top-8 setting despite both having about 2.9B parameters, suggesting that distributing routed residual capacity across the pretrained decoder is more effective than concentrating many experts in fewer MoE layers. By contrast, adding more experts within each MoE layer is not consistently beneficial. Larger expert/activation configurations often underperform the default 8-expert/Top-2 setting unless the parameter count is substantially increased. Increasing routing Top-k improves performance when the number of experts is fixed, but it also increases per-token computation; therefore, we use Top-2 as a practical balance between adaptation quality and sparse activation efficiency.

Configuration	Params	Avg. COMET	Avg. BLEU
MoE layer interval			
Every 1 layer	2.9B	**80.57**	**24.43**
Every 2 layers	1.8B	78.42	23.37
Every 4 layers	1.3B	78.42	22.24
Every 8 layers	1.0B	77.09	21.10
Expert / activation configuration			
8 experts, Top-2	1.3B	78.42	22.24
16 experts, Top-4	1.8B	77.92	21.90
32 experts, Top-8	2.9B	78.14	22.09
64 experts, Top-8	5.0B	**78.61**	**22.59**
Routing Top-k / residual grouping			
8 experts, Top-1	1.3B	77.47	21.44
8 experts, Top-2	1.3B	78.42	22.24
8 experts, Top-4	1.3B	78.48	22.28
8 experts, Top-6	1.3B	78.31	22.15
8 experts, Top-8	1.3B	**78.80**	**22.53**

Table 6: Sensitivity to MoE architecture configurations. We study MoE layer placement, expert/activation scaling, and routing Top-k. Full per-language results are provided in Appendix A.7.

5 Related Work

Mixture of Experts. Mixture-of-Experts (MoE) models scale parameter capacity by replacing dense FFN layers with multiple experts and activating only a sparse subset for each input (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). This design has been widely used in large-scale language models to improve capacity without proportionally increasing computation (Dai et al., 2024; Muennighoff et al., 2025). MoE also enables expert specialization, where different experts can capture different domains, languages, or input patterns (Jacobs et al., 1991; Ma et al., 2018; Gupta et al., 2022). In contrast to training MoE models from scratch, our work studies MoE upcycling and focuses on initializing routed experts with structured diversity grounded in pretrained dense weights.

MoE upcycling. MoE upcycling expands a pretrained dense model into a sparse MoE model by reusing dense FFN weights for expert initialization (Komatsuzaki et al., 2022; He et al., 2024). Early upcycling methods typically duplicate the dense FFN into routed experts, while recent variants introduce perturbation-based initialization, such as noise injection, partial re-initialization, or drop-based construction, to reduce expert homogeneity (Yang et al., 2024; Muennighoff et al., 2025; Nakamura et al., 2025). Branch-Train-MiX (Sukhbaatar et al., 2024) instead trains multiple dense branches and mixes their FFN modules into a unified MoE model. However, duplicated experts lack diversity, whereas unstructured perturbations can disrupt pretrained knowledge, especially when large-scale continued pretraining is unavailable. SPRI addresses this trade-off by initializing routed experts with SVD-partitioned residuals from pretrained FFN weights, yielding structured expert diversity grounded in the dense checkpoint.

6 Conclusion

We study MoE upcycling under data-constrained adaptation, where conventional initialization strategies struggle to balance pretrained knowledge preservation and routed expert diversity. We propose SPRI, an SVD-partitioned residual initialization method that constructs routed experts from structured spectral residuals of pretrained FFN weights. By grounding expert diversity in the pretrained weight structure, SPRI leverages the pretrained weight structure while introducing diversity among routed experts.

Experiments on the CoVoST2 S2TT dataset show that SPRI outperforms dense fine-tuning and conventional MoE upcycling baselines in both BLEU and COMET. Our ablations further show that the SVD-based residual construction, two-stage training, and MoE architecture choices all affect the effectiveness of data-constrained upcycling. Overall, our results suggest that controlled, weight-grounded residual diversity is an effective principle for MoE upcycling when large-scale continued pretraining is unavailable.

7 Limitations

This work has several limitations.

- **Limited evaluation scope.** Our experiments focus on CoVoST 2 English-to-X speech-to-text translation. Although this setting is suitable for studying data-constrained MoE upcycling with multiple target directions, it does not fully establish whether the same trends hold for other adaptation tasks or domains. Extending SPRI to broader speech-language and text-only adaptation settings remains future work.

- **Limited model and training scale.** We mainly evaluate SPRI on Qwen3-ASR-0.6B under a limited-data adaptation budget. While this setting matches our motivation, it may make architecture-level observations, such as the effects of MoE layer interval, expert count, and routing Top-k, sensitive to the training scale. Therefore, these findings should be interpreted as empirical guidance for data-constrained upcycling rather than general scaling rules for MoE models. Extending SPRI to larger dense checkpoints, such as 7B- and 13B-scale models, and larger continued-training budgets remains an important direction for future work.

- **Limited joint hyperparameter search.** To ensure fair comparisons, our main experiments use a fixed MoE configuration, and our ablations vary one factor at a time. While this isolates the effect of each design choice, it does not exhaustively explore interactions among MoE architecture, residual placement, routing Top-k, and training schedules. Some ablation results therefore suggest that stronger SPRI configurations may exist under joint tuning. A more comprehensive architecture and training search is left for future work.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024a. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

DeepSeek-AI. 2024. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

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.

Jiatao Gu, Hany Hassan Awadalla, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Shashank Gupta, Subhabrata Mukherjee, Krishan Subudhi, Eduardo Gonzalez, Damien Jose, Ahmed H Awadallah, and Jianfeng Gao. 2022. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arXiv:2204.07689*.

Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath

Aithal, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Upcycling large language models into mixture of experts. *arXiv preprint arXiv:2410.07524*.

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.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2022. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Thang Luong, and Orhan Firat. 2021. Exploring routing strategies for multilingual mixture-of-experts models.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. 2025. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pages 62061–62121.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. 2025. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. *arXiv preprint arXiv:2502.19261*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

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.

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*.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, volume 2021, pages 2247–2251.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Zhifang Guo Yongqi Wang Pei Zhang Xinyu Zhang Zishan Guo Hongkun Hao Yu Xi Baosong Yang Jin Xu Jingren Zhou Junyang Lin Xian Shi, Xiong Wang. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Appendix

A.1 Diversity and KL Divergenc

We analyze two aspects of the upcycled MoE models: the diversity among initialized routed experts and the deviation from the pretrained dense model during training.

Inter-expert diversity. We compute diversity only from the initialized routed-expert down-projection weights. For each MoE layer, we take the 8 matrices $\{\mathbf{W}_{\ell,i}^{(\mathrm{Down})}\}_{i=1}^{8}$ and compute the average pairwise cosine similarity after vectorizing each matrix. The diversity score of layer ℓ is defined as one minus this average similarity:

$$\mathrm{Div}_\ell = 1 - \operatorname*{Avg}_{i<j} \cos\Big(\mathrm{vec}(\mathbf{W}_{\ell,i}^{(\mathrm{Down})}), \mathrm{vec}(\mathbf{W}_{\ell,j}^{(\mathrm{Down})})\Big)$$

We then average Div_ℓ over all MoE layers. The routed gate and up projections are not used in this metric. For SPRI, experts assigned to the same SVD group share the same spectral residual up to a small tie-breaking perturbation.

KL divergence. We measure how much each upcycled model deviates from the pretrained dense model using a fixed validation subset. Specifically, we sample 1000 validation examples with shuffle seed 42 and evaluate them as 10 batches of size 1. We first run the original dense model on this subset and cache its per-token output distributions as the reference. At each evaluated training step, we run each upcycled model on the same subset under teacher forcing and compute the token-level KL divergence to the cached dense reference. KL divergence is computed only on target positions included in the training loss, excluding labels equal to -100.

A.2 Experimental Details

Data preprocessing and loading. During data loading, we shuffle the training data at the dataset level to avoid language-direction clustering in early training steps. This ensures that initial optimization batches already contain samples from different translation directions, which helps stabilize router learning in MoE models. Notably, because Common Voice 4 is no longer publicly accessible, we use an externally processed version of the dataset instead of reconstructing CoVoST 2 from Common Voice 4 with the official CoVoST 2 preprocessing scripts.

Training data. We train on English-source speech-to-text translation pairs assembled from two publicly available corpora: CoVoST 2 and Europarl-ST. After filtering utterances longer than 15 s or whose tokenized prompts exceeded 512 tokens, totaling 6933.9 hours of English speech. CoVoST 2 contributes 6451.9 h and covers 15 target languages, with an essentially uniform allocation of 430.13h $\times$ 15 per direction. Europarl-ST provides an additional 482.0 h drawn from European Parliament proceedings, covering 8 target languages with translation hours of de: 62.78, pt: 61.44, fr: 61.40, es: 61.35, nl: 60.82, pl: 60.02, it: 59.27, ro: 54.92.

Model Architecture. We experimented on Qwen3-ASR-0.6B. The acoustic encoder contains 18 Transformer layers, and the language decoder contains 28 Transformer layers with a hidden size of 1024, FFN intermediate size 3072, and 16 attention heads.ention heads.

Tools. We fine-tune all models with the Hugging Face Transformers library[5], using the default configuration of Qwen3-ASR-0.6B unless otherwise specified[6]. We used GPT-5.5 solely for writing refinement. All research ideas, experimental design, experiments, and analyses were conducted by the authors.

A.3 Full Results for Routed Down-Projection Initialization

Table 8 reports the full per-language COMET and BLEU results for the routed down-projection initialization ablation. All variants use the same SPRI framework, including the shared expert, copied routed gate/up projections, MoE architecture, and two-stage training strategy. They differ only in how the routed down projections are initialized. This ablation evaluates whether the improvement of SPRI comes from the proposed SVD-partitioned residual construction rather than from the surrounding MoE architecture or training strategy.

Instead of removing the dense-copied expert, the w/o shared expert variant treats it as a regular routed expert and lets the router decide when to activate it.

[5] https://github.com/huggingface/transformers
[6] https://huggingface.co/Qwen/Qwen3-ASR-0.6B

Hyperparameter	Value
Audio encoder layers / dim. / heads	18 / 896 / 14
Decoder layers / dim. / heads (Q/KV)	28 / 1024 / 16/8
FFN intermediate dim.	3072
Vocabulary size	151936

Table 7: Architecture of Qwen3-ASR-0.6B.

	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
Model	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Down-projection initialization																
Copy	77.04	12.24	75.59	8.92	85.94	30.23	68.99	14.71	66.63	10.40	68.86	11.06	65.64	11.00	80.42	28.15
Noise	77.07	12.25	75.50	8.74	85.87	30.23	68.77	14.58	66.59	10.39	68.83	10.99	65.52	11.07	80.42	28.19
Drop	75.66	11.54	74.96	8.60	85.19	29.01	68.17	14.51	65.95	10.18	67.45	10.54	64.69	10.63	79.04	27.02
SVD-Partitioned	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
Shared expert																
w/ shared expert	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
w/o shared expert	77.93	13.83	79.49	11.52	86.08	30.80	74.06	19.13	73.99	14.20	71.83	13.75	72.78	14.14	80.67	29.16

	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
Model	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Down-projection initialization																
Copy	84.06	41.79	68.68	4.80	76.12	26.56	68.53	17.21	75.50	26.48	79.78	18.04	85.63	32.99	75.16	19.64
Noise	84.06	41.82	68.59	4.73	76.09	26.34	68.45	16.96	75.42	26.55	79.80	18.03	85.63	32.96	75.11	19.59
Drop	83.42	40.37	68.76	5.06	74.04	25.13	68.64	17.13	74.36	25.84	78.63	17.15	84.57	31.45	74.23	18.94
SVD-Partitioned	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
Shared expert																
w/ shared expert	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
w/o shared expert	84.37	43.04	76.36	7.94	76.96	28.52	76.53	24.06	77.07	30.42	80.06	19.03	85.93	34.14	78.28	22.24

Table 8: Full per-language results for the routed down-projection initialization ablation on CoVoST2 English-to-X speech-to-text translation. All variants use the same SPRI framework and differ only in how routed down projections are initialized. The last columns report the average COMET and BLEU across all translation directions.

A.4 Full Results for Parameter-Matched Dense Comparison

Table 9 reports the full per-language COMET and BLEU results for the parameter-matched comparison. We compare SPRI-3B with Qwen2.5-3B before and after fine-tuning. SPRI-3B is obtained by upcycling a smaller pretrained model into a sparse MoE model with a comparable total parameter scale. The results show that SPRI-3B consistently improves the average performance over both dense baselines, supporting MoE upcycling as an effective scaling strategy under data-constrained S2TT adaptation.

A.5 Detailed Results for Unfreezing Ratio Sensitivity

Table 10 reports the full per-language COMET and BLEU results for the freezing-ratio ablation in Section **??**. The freeze value denotes the fraction of training steps during which the routed-expert gate and up projections are kept frozen.

A.6 Full Results for Residual Projection Placement

Table 11 reports the full per-language BLEU results for the residual projection placement ablation. All variants use the same training data, MoE configuration, optimizer settings, and training budget; they only differ in whether the SVD-partitioned residual initialization is applied to the up, gate, or down projection. The average BLEU values correspond to the results summarized in Table 5.

A.7 Full Results for MoE Architecture Sensitivity

Table 12 reports the full per-language BLEU results for the MoE architecture sensitivity experiments. We evaluate the effect of MoE layer placement, expert/activation configuration, and grouping-related hyperparameters. All variants use the same training data, optimizer settings, and training budget unless otherwise specified.

Model	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
SPRI 3B	80.17	15.22	83.27	14.92	86.39	31.61	77.90	22.53	78.03	17.91	76.19	16.02	76.97	17.95	81.27	29.46
Qwen2.5 3B	71.72	6.88	50.79	0.06	86.35	30.25	61.09	5.40	54.93	1.31	64.61	7.08	54.12	2.66	81.78	29.16
Qwen2.5 3B FT	78.01	11.52	72.20	6.34	86.59	30.69	73.21	16.10	69.81	10.84	72.29	11.51	69.51	11.36	82.23	29.94

Model	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
SPRI 3B	84.45	43.09	80.04	11.33	79.67	30.20	78.10	29.11	79.19	32.95	80.78	19.93	86.19	34.30	80.57	24.43
Qwen2.5 3B	85.29	41.99	46.48	0.26	76.32	21.33	52.20	1.24	72.14	18.05	80.41	18.82	82.77	23.39	68.07	13.86
Qwen2.5 3B FT	85.31	44.83	66.85	3.86	79.92	27.91	68.30	15.54	78.15	28.73	81.08	19.91	85.34	30.12	76.59	19.95

Table 9: Full per-language results for the parameter-matched dense comparison on CoVoST2 English-to-X speech-to-text translation. SPRI-3B is compared with `Qwen2.5-Omni-3B` and its fine-tuned variant at a comparable total parameter scale.

Freeze	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
0	78.51	13.58	78.93	11.02	86.24	30.69	72.58	17.66	71.05	12.98	72.26	13.24	70.71	13.62	80.98	29.04
0.05	78.79	13.75	80.21	11.95	86.29	30.79	74.12	19.12	73.44	14.30	73.54	13.93	72.83	14.79	81.00	29.16
0.1	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
0.2	78.74	13.78	79.57	11.45	86.29	30.93	73.83	18.63	72.84	13.85	72.84	13.53	71.50	14.05	81.02	29.06
0.5	78.59	13.52	79.45	11.42	86.31	30.83	73.15	17.90	71.94	13.57	72.58	13.46	71.50	13.83	80.97	29.07

Freeze	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
0	84.56	42.85	73.77	7.11	77.72	28.15	73.36	22.22	77.10	29.30	80.41	18.80	86.12	33.89	77.62	21.61
0.05	84.62	42.92	76.36	8.54	77.96	28.69	75.73	24.99	77.65	30.39	80.37	19.04	86.14	33.98	78.60	22.42
0.1	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
0.2	84.59	42.99	75.00	7.57	77.91	28.24	74.41	23.31	77.70	30.29	80.44	18.90	86.20	34.09	78.19	22.04
0.5	84.58	42.99	74.62	7.46	77.89	28.23	74.05	23.04	77.58	30.09	80.45	19.12	86.19	34.04	77.99	21.90

Table 10: Full per-language results for the freezing-ratio ablation on CoVoST2 English-to-X speech-to-text translation. We report COMET and BLEU for each target language and their average across all directions. The freeze value denotes the fraction of training steps during which the routed-expert gate and up projections are kept frozen.

Model	Param	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
SPRI - Up	1.3B	78.42	13.48	79.02	11.15	86.20	30.64	73.13	18.18	70.97	13.00	72.00	13.05	70.41	13.80	80.98	29.16
SPRI - Down	1.3B	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
SPRI - Gate	1.3B	78.72	13.70	80.92	12.50	86.25	30.79	73.91	18.85	73.58	14.51	74.04	14.30	72.94	14.67	80.98	29.12

Model	Param	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
SPRI - Up	1.3B	84.52	42.88	73.89	7.18	77.79	28.34	74.14	23.07	77.10	29.35	80.35	18.81	86.12	33.89	77.67	21.73
SPRI - Down	1.3B	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
SPRI - Gate	1.3B	84.56	42.99	76.83	8.61	78.10	28.70	75.58	24.85	78.11	30.92	80.44	19.09	86.17	34.07	78.74	22.51

Table 11: Full per-language results for the SVD residual placement ablation on CoVoST2 English-to-X speech-to-text translation. We compare applying SVD-partitioned residuals to the up, down, and gate projections while keeping the remaining MoE architecture and training setup fixed. We report COMET and BLEU for each target language and their average across all directions.

	Tr		Ta		Ja		Sl		Lv		Fa		Et		De	
Model	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Every 1 layer	80.17	15.22	83.27	14.92	86.39	31.61	77.90	22.53	78.03	17.91	76.19	16.02	76.97	17.95	81.27	29.46
Every 2 layers	78.85	14.28	79.80	13.26	86.28	31.17	74.25	20.79	73.45	16.19	73.08	15.02	71.90	16.20	81.03	29.23
Every 4 layers	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
Every 8 layers	77.88	13.05	78.50	10.56	86.20	30.61	72.15	17.06	71.15	12.94	70.67	12.31	68.62	12.27	80.87	28.73
8 experts, Top-2	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
16 experts, Top-4	78.56	13.50	79.45	11.40	86.30	30.72	73.28	18.33	72.23	13.72	72.11	13.12	71.18	14.03	81.03	29.11
32 experts, Top-8	78.60	13.65	79.56	11.51	86.34	31.13	73.48	18.36	72.73	13.94	72.74	13.55	71.62	14.21	81.04	29.14
64 experts, Top-8	79.07	14.01	80.01	11.80	86.30	31.30	74.38	19.49	73.64	14.78	73.10	13.87	72.52	14.99	81.14	29.18
1 expert, Top-1	77.96	13.10	78.26	10.41	86.15	30.82	72.31	17.44	71.72	13.19	71.34	12.62	70.26	13.02	80.88	28.90
2 experts, Top-2	78.85	13.83	79.80	11.52	86.28	30.80	74.25	19.13	73.45	14.20	73.08	13.75	71.90	14.14	81.03	29.16
4 experts, Top-4	78.61	13.59	80.04	11.82	86.26	30.78	73.95	18.86	73.48	14.45	73.42	13.78	71.98	14.20	80.94	29.07
6 experts, Top-6	78.33	13.51	79.91	11.71	86.21	30.75	73.66	18.53	73.11	14.01	73.05	13.54	71.67	14.19	80.95	29.07
8 experts, Top-8	78.54	13.57	80.01	11.71	86.21	30.72	74.83	19.46	74.44	14.75	73.65	14.03	73.58	15.21	81.06	29.18

	Zh		mn		Sv		Cy		Ca		Ar		Id		Avg	
Model	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Every 1 layer	84.45	43.09	80.04	11.33	79.67	30.20	78.10	29.11	79.19	32.95	80.78	19.93	86.19	34.30	80.57	24.43
Every 2 layers	84.57	43.07	75.57	9.72	77.98	29.22	75.05	27.51	77.81	31.72	80.46	19.10	86.26	34.10	78.42	23.37
Every 4 layers	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
Every 8 layers	84.49	42.58	72.63	6.49	77.25	27.60	73.04	21.72	76.54	28.15	80.25	18.74	86.07	33.76	77.09	21.10
8 experts, Top-2	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
16 experts, Top-4	84.60	42.93	74.13	7.29	77.95	28.42	73.96	22.99	77.27	29.57	80.49	19.02	86.25	34.30	77.92	21.90
32 experts, Top-8	84.59	42.86	74.71	7.57	77.92	28.59	74.29	23.38	77.78	30.42	80.46	19.14	86.24	33.92	78.14	22.09
64 experts, Top-8	84.56	43.09	75.95	8.45	78.49	29.19	75.19	24.62	77.90	30.53	80.57	19.30	86.31	34.30	78.61	22.59
1 expert, Top-1	84.51	42.72	73.46	6.78	77.21	27.60	74.79	23.61	76.97	29.09	80.21	18.65	86.05	33.73	77.47	21.44
2 experts, Top-2	84.57	43.04	75.57	7.94	77.98	28.52	75.05	24.06	77.81	30.42	80.46	19.03	86.26	34.14	78.42	22.24
4 experts, Top-4	84.56	42.86	75.91	8.09	77.79	28.41	76.20	25.53	77.51	30.02	80.41	19.00	86.13	33.81	78.48	22.28
6 experts, Top-6	84.52	42.91	75.40	7.79	77.71	28.15	76.08	25.44	77.50	29.98	80.34	18.69	86.16	34.02	78.31	22.15
8 experts, Top-8	84.58	43.00	75.83	8.15	77.92	28.41	76.65	26.13	77.93	30.47	80.52	19.30	86.22	33.86	78.80	22.53

Table 12: Full per-language results for the MoE architecture sensitivity study on CoVoST2 English-to-X speech-to-text translation. We report COMET and BLEU for each target language and their average across all directions. The evaluated configurations include MoE layer interval, expert/activation scaling, and routing Top-k under SPRI-based upcycling.