

MeRP: Efficient Multi-Expert Routing Prompt Encoding for Multilingual Translation with LLMs

Fengzhao Fan^{1,2}, Zhenhan Wang^{1,2*}, Shengxiang Gao^{1,2}, Cunli Mao^{1,2}, Zhengtao Yu^{1,2}

¹Faculty of Information Engineering and Automation,

Kunming University of Science and Technology, Kunming, China

²Yunnan Key Laboratory of Artificial Intelligence, Kunming, China

fengzhao_fan@163.com, 18640248002@163.com, gaoshengxiang.yn@hotmail.com,

maocunli@163.com, ztyu@hotmail.com

Abstract

Large Language Models (LLMs) enable a single model to support translation across multiple language directions simultaneously, eliminating the need to train a separate model for each direction and substantially improving the performance of multilingual machine translation. However, during multilingual translation instruction tuning, parameter interference among different languages significantly limits the improvement of translation performance and may even cause performance degradation for some languages. To address this issue, we propose MeRP, a multi-expert routing prompt encoding strategy for parameter-efficient fine-tuning (PEFT) in multilingual machine translation tasks. This method encodes representations of source languages through an attention mechanism and dynamically selects expert groups according to translation directions, enabling both the sharing and disentanglement of multilingual semantic representations. As a result, it preserves LLM generalization while reducing training time and improving translation performance. We conduct experiments on five language pairs. The results show that, compared with the baseline models, our method updates only 0.06% of the total model parameters during training while achieving better multilingual translation performance. Notably, our MeRP reduces training time by 46.75% and achieves average gains of +0.62 BLEU points and +0.34 COMET-22 points over the arithmetic mean of the four PEFT baselines. Code is available at <https://github.com/wangzhenhan93/MeRP>.

1 Introduction

LLMs have recently defined the new state-of-the-art in MT, demonstrating remarkable cross-lingual generalization capabilities (Cui et al., 2025; Zhu et al., 2024b; Lai et al., 2023). By leveraging massive pre-training data, LLMs effectively capture

*Corresponding author.

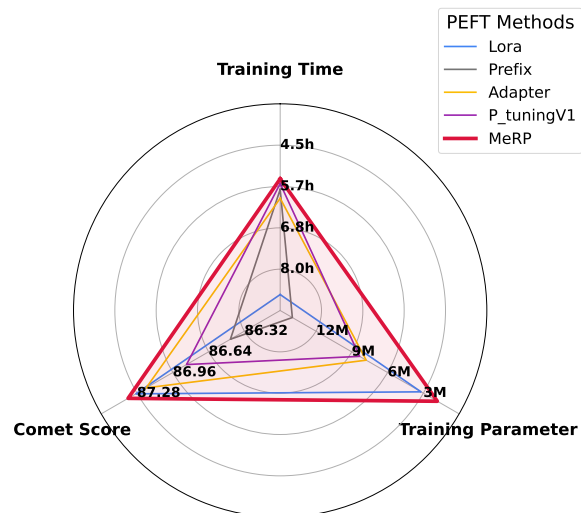


Figure 1: PEFT comparison on Gemma-2-2B-It. Radar chart evaluates five methods across three normalized metrics (lower training time, fewer parameters, higher COMET-22), where outer values indicate better performance. The solid crimson line (MeRP) shows the best overall performance.

universal linguistic patterns, demonstrating exceptional multilingual proficiency. This inherent capability lays a solid foundation for multilingual machine translation (MMT), making the adaptation of general-purpose models a pivotal research direction (Dong et al., 2025; Zhu et al., 2024a) that promises to unify diverse translation directions within a single representation space.

To adapt these massive models for MMT, PEFT has emerged as the standard paradigm. Techniques such as LoRA (Hu et al., 2022), Prefix-Tuning (Li and Liang, 2021), and Adapters (Alves et al., 2023) introduce only a small number of trainable parameters. These methods circumvent the instability and excessive resource demands of full-parameter fine-tuning, allowing LLMs to be effectively steered toward translation tasks.

However, applying standard PEFT to MMT introduces a fundamental challenge: cross-lingual

parameter interference arising from linguistic heterogeneity. Standard PEFT methods typically employ a globally shared set of trainable parameters across all languages. Given the vast intrinsic differences between languages, forcing linguistically diverse directions (e.g., English→Chinese and English→German) to share a unified, low-rank parameter space creates a competition for model capacity. This phenomenon, often referred to as the “curse of multilinguality,” has been associated with conflicting gradient updates, where dominant high-resource languages may overwrite representations needed by low-resource languages. Consequently, the model suffers from capacity dilution, leading to suboptimal performance and a failure to capture language-specific nuances.

To address this issue, we propose MeRP, a multi-expert routing prompt encoding strategy for PEFT in MMT tasks. Under fully frozen LLM backbones, MeRP utilizes a lightweight external encoder to encode source-language representations through an attention mechanism and dynamically select expert groups according to translation directions. The encoder first extracts reusable cross-directional translation priors through a Shared Cross-Attention Trunk, subsequently allocates language-direction-specific residual information via a routing expert pool, and finally stabilizes expert preferences using a soft anchoring mechanism. This design achieves an effective separation of universal translation priors and direction-specific adaptation information within a single compact module, providing a parameter-efficient and interpretable solution for multilingual translation.

We conducted experiments on five language pairs with three types of frozen backbone models. The results show that MeRP requires only 0.06% of the total model parameters for training while achieving better multilingual translation performance. Notably, MeRP accelerates training by 46.75% and yields an average improvement of +0.62 BLEU points and +0.34 COMET-22 score.

Our contributions are as follows:

- We introduce MeRP, a multi-expert routing prompt encoding strategy that combines a source-conditioned Shared Cross-Attention Trunk and direction-dependent routing experts to generate direction-aware prompts with frozen LLM backbones.
- MeRP encodes source-language representations through an attention mechanism and dy-

namically selects expert groups according to translation directions, enabling both the sharing and disentanglement of multilingual semantic representations.

- MeRP achieves strong cross-lingual generalization with minimal parameter overhead, maintaining a balanced trade-off between training efficiency and translation quality across multiple LLMs and PEFT methods.

2 Related Work

2.1 Multilingual Machine Translation with Large Language Models

Large Language Models (LLMs) acquire general-purpose multilingual representations through large-scale cross-lingual pretraining and have become a fundamental paradigm for MMT. Recent studies show that instruction-tuned multilingual Large Language Models exhibit strong zero-shot and few-shot machine translation capabilities, demonstrating effective cross-lingual transfer across a wide range of languages, particularly in low-resource settings (Qin et al., 2025). Empirical studies further show that state-of-the-art LLMs, including GPT-4 and ChatGPT, can achieve performance comparable to or even surpassing conventional supervised machine translation systems on certain translation directions. However, their effectiveness is still influenced by language typological variation, data distribution imbalance, and differences in translation difficulty across directions (Zhu et al., 2024b).

Despite their strong multilingual capabilities, most existing approaches to multilingual adaptation rely on unified or highly shared parameterizations. While such designs facilitate cross-lingual knowledge transfer, they also tend to introduce dominance effects from high-resource languages, leading to cross-lingual semantic interference and negative transfer (Zhu et al., 2021; Shaham et al., 2023; Zheng et al., 2025; Dong et al., 2025). Consequently, enabling lightweight and direction-specific adaptation across translation directions while preserving the general multilingual competence of LLMs remains a core challenge.

2.2 Parameter-Efficient Fine-Tuning Methods for Large Language Models

To mitigate cross-lingual interference under frozen-backbone settings, PEFT methods adapt pretrained models by training only a small number of additional parameters. Representative approaches

include Adapter (Alves et al., 2023), Prefix-Tuning (Li and Liang, 2021), and LoRA (Hu et al., 2022). Prior studies also show that prompt- and prefix-based adaptation can improve translation performance in LLM-based systems (Zhang et al., 2023; Gao et al., 2024; Yamada, 2023).

In multilingual machine translation, however, standard PEFT methods usually share one trainable parameter space across languages, which can intensify competition over limited adaptation capacity. To alleviate this issue, recent work explores language-aware routing and sparse parameter activation for multilingual adaptation (Zhu et al., 2024a; Zhao et al., 2024), while systematic studies further show that PEFT architectures vary substantially across low-resource multilingual settings (Su et al., 2024).

MoE models provide a related but distinct line of research by improving parameter utilization through dynamic expert routing (Fedus et al., 2022). Yet they are mainly designed for scalable conditional computation rather than parameter-efficient, direction-aware adaptation for frozen multilingual LLMs. Therefore, designing a lightweight routing mechanism that preserves shared translation priors while modeling direction-specific residual information remains an open problem.

3 Methodology

We propose **MeRP**, a multi-expert routing prompt encoding strategy for PEFT in multilingual machine translation. Its core idea is to leverage lightweight modules to extract cross-directional translation priors and model language-direction-specific residuals, thereby mitigating parameter interference in multilingual fine-tuning.

The core architecture of MeRP consists of the Shared Cross-Attention Trunk for encoding source-language representations and the Multi-expert Routed Module, as illustrated in Figure 2. The Shared Cross-Attention Trunk extracts shared translation priors from the source representations, while the Multi-expert Routed Module comprises the Direction Route and Expert Pool, which compute expert weights based on the translation direction and generate a Direction-specific Residual Prompt, with the soft anchoring mechanism stabilizing expert preferences. The Prompt Fusion Module then combines the shared representation with the Direction-specific Residual Prompt and projects the fused representation back into the em-

bedding space of the backbone, where it is injected into the frozen LLM as dynamic prefixes. Unlike LoRA and Adapters, which adapt the model by modifying or augmenting internal backbone layers, MeRP performs adaptation through an external prompt encoder, where the low-rank experts generate prompt representations rather than update the backbone parameters.

3.1 Shared Cross-Attention Trunk

The Shared Cross-Attention Trunk uses learnable prompt query vectors to interact with source sequence representations through cross-attention, aiming to extract common prior knowledge across translation directions from the source sequence. The representations generated by this module serve as the foundation for subsequent direction-adaptive experts modeling and prompt injection.

Given the source-side input embedding $x \in R^{B \times S \times D}$, where B denotes the batch size, S denotes the source sequence length, and D denotes the embedding dimension of the frozen backbone, the method first maps the input into a bottleneck space via a linear down-projection:

$$h_0 = \text{LayerNorm}(x)W_{\text{down}}, W_{\text{down}} \in R^{D \times d_t} \quad (1)$$

where W_{down} is the down-projection matrix and d_t is the representation dimension of the shared trunk. Let h_0 act as the Key (K) and Value (V) in the cross-attention mechanism, such that $K = V = h_0$. Within the shared trunk, a set of globally shared learnable prompt query vectors $Q \in R^{1 \times P \times d_t}$ is maintained, where P represents the prompt length. Unlike the self-attention mechanism, Q is not mapped from the input sequence but instead serves as a set of learnable parameters to read common translation priors from the source sequence. Layer normalization is applied to the queries and the key-value pairs respectively:

$$\tilde{Q} = \text{LayerNorm}_Q(Q) \quad (2)$$

$$\tilde{K} = \text{LayerNorm}_{KV}(K), \tilde{V} = \tilde{K} \quad (3)$$

Multi-head cross-attention is used to calculate alignment between the queries and the source representations:

$$S_h = \frac{(\tilde{Q}W_Q^h)(\tilde{K}W_K^h)^\top}{\sqrt{d_h}} + M \quad (4)$$

$$\mathbf{A}_h = \text{softmax}(S_h)(\tilde{V}W_V^h) \quad (5)$$

$$\text{AttnOut} = \text{Concat}(\mathbf{A}_1, \dots, \mathbf{A}_H) \quad (6)$$

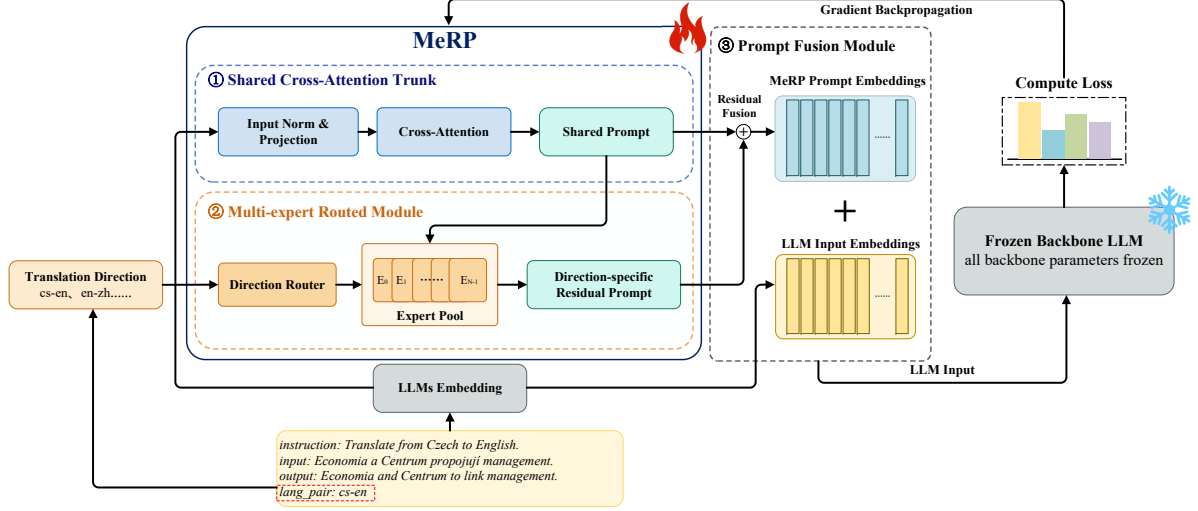


Figure 2: Overview of MeRP. Source embeddings are encoded through the Shared Cross-Attention Trunk, while expert groups are dynamically selected by translation direction for shared yet disentangled multilingual representations.

where $d_h = d_t/H$, H is the number of heads, and M is a mask matrix used to prevent invalid positions, such as padding, from participating in weight allocation. To ensure training stability, the attention output is added to the original queries to obtain the intermediate representation:

$$Z = Q + \text{Dropout}(\text{AttnOut}) \quad (7)$$

On this basis, a non-linear mapping is performed through a Feed-Forward Network (FFN). Denoting the weight parameters of the FFN as W_{in} and W_{out} , the final output of the shared trunk is:

$$h_{\text{shared}} = Z + \text{Dropout}(\phi(\text{LayerNorm}_{\text{ffn}}(Z)W_{\text{in}})W_{\text{out}}) \quad (8)$$

where $\phi(\cdot)$ denotes the GELU activation and $h_{\text{shared}} \in R^{B \times P \times d_t}$. This representation is shared across all translation directions and is responsible for modeling shared translation priors.

3.2 Direction-Adaptive and Soft Anchoring

Building upon the shared trunk output h_{shared} , this method further introduces direction-adaptive experts to model the residual specialization of different translation directions. Assuming there are N translation directions in total, for the i -th direction, the module first obtains its direction embedding:

$$e_i \in R^{d_{\text{dir}}} \quad (9)$$

where d_{dir} is the dimension of the direction embedding. The routing distribution over K experts

is then computed via a linear gating layer:

$$g_i = \text{softmax}(e_i W_g + b_g), g_i \in R^K \quad (10)$$

where W_g and b_g denote the gating projection matrix and the bias term, respectively; K represents the number of experts, and $g_{i,k}$ denotes the weight assigned to the k -th expert by direction i . Each expert employs a low-rank residual mapping. For the k -th expert, let $W_k^{\text{down}} \in R^{d_t \times r}$ and $W_k^{\text{up}} \in R^{r \times d_t}$ be the down-projection and up-projection matrices, respectively, where r is the low-rank dimension.

The expert mapping is defined as:

$$E_k(h) = \phi(h W_k^{\text{down}}) W_k^{\text{up}} \quad (11)$$

Thus, the private residual representation for direction i can be expressed as:

$$h_{\text{priv}}^{(i)} = \sum_{k=1}^K g_{i,k} E_k(h_{\text{shared}}) \quad (12)$$

To enhance the stability and interpretability of the routing structure, we further introduce a soft anchoring mechanism. For each direction i , a primary expert a_i is pre-specified, and a small set of overflow experts $\mathcal{O}$ is reserved as controlled shared capacity. During training, rather than forcing the routing into a one-hot distribution, we construct a soft target distribution q_i for each direction. Let α be the target probability for the primary expert and β be the target probability for the overflow experts.

$$q_{i,k} = \begin{cases} \alpha, & k = a_i, \\ \beta, & k \in \mathcal{O}, \\ \frac{1-\alpha-|\mathcal{O}|\beta}{K-1-|\mathcal{O}|}, & \text{otherwise.} \end{cases} \quad (13)$$

where $|\mathcal{O}|$ denotes the number of overflow experts. This distribution allocates the majority of the probability to the primary expert, a small portion to the overflow experts, and the remainder equally among the other experts.

To ensure the dominance of the primary expert and suppress unnecessary routing drift, this module utilizes a combination of soft cross-entropy and margin regularization to form the anchoring loss. Let the routing logits for direction i be:

$$z_i = W_g e_i + b_g \quad (14)$$

The anchoring loss is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{anchor}} = & -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K q_{i,k} \log g_{i,k} \\ & + \lambda_m \frac{1}{N} \sum_{i=1}^N \max\left(0, m - (z_{i,a_i} \right. \\ & \left. - \max_{k \neq a_i} z_{i,k})\right) \end{aligned} \quad (15)$$

where m denotes the margin and λ_m denotes the weight of the margin term.

To further discourage routing collapse, we introduce a contrastive routing loss that operates on normalized routing distributions. For each translation direction i , we first normalize its routing probability vector:

$$\hat{g}_i = \frac{g_i}{\|g_i\|_2}, \quad (16)$$

where $\hat{g}_i$ denotes the L2-normalized routing distribution of direction i . Based on this normalization, the similarity between two directions is measured by their cosine similarity, and the contrastive routing loss is defined as:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{N(N-1)} \sum_{i \neq j} \hat{g}_i^\top \hat{g}_j. \quad (17)$$

This term is weighted by λ_c and added to the overall objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{anchor}} + \lambda_c \mathcal{L}_{\text{contrast}}. \quad (18)$$

Here, $\mathcal{L}_{\text{main}}$ denotes the translation loss.

3.3 Representation Generation and Injection

After obtaining the shared representation h_{shared} and the direction-private residual $h_{\text{priv}}^{(i)}$, this method fuses the two to derive the direction-aware prompt representation:

$$h_{\text{out}}^{(i)} = h_{\text{shared}} + h_{\text{priv}}^{(i)} \quad (19)$$

Upon obtaining $h_{\text{out}}^{(i)}$, an up-projection matrix $W_{\text{up}} \in R^{d_t \times D}$ maps it back to the backbone embedding space. After output normalization, the final dynamic prefix is obtained as

$$\begin{aligned} p^{(i)} &= s \cdot \text{LayerNorm}_{\text{out}}(h_{\text{out}}^{(i)} W_{\text{up}}) \\ p^{(i)} &\in R^{B \times P \times D} \end{aligned} \quad (20)$$

where s is initialized to 1.1×10^{-3} to stabilize early training, and W_{up} is initialized near zero to minimize early perturbation to the frozen backbone. These prefix representations are injected into the frozen LLM backbone to activate latent knowledge relevant to the target translation direction without modifying backbone parameters.

4 Experiments

We conducted multi-directional translation experiments on three frozen Large Language Models, Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B, and compared our approach with representative PEFT methods to systematically evaluate MeRP in terms of parameter efficiency, translation quality, and hardware feasibility.

4.1 Datasets

To evaluate MeRP on multilingual translation, we construct the training data from publicly available parallel corpora, mainly NiuTrans/LMT-60-sft-data¹ (Yingfeng Luo, 2025) and ALT². NiuTrans/LMT-60-sft-data includes multilingual translation resources collected from benchmarks such as FLORES-200 devset, NTREX-128 (Fiedermann et al., 2022), SMol, WMT14-23, and IWSLT17-24, while ALT (Asian Language Treebank) provides additional parallel data for several Asian languages. Our training setup is English-centric and covers Czech, Vietnamese, Russian, German, and Chinese, thereby spanning high-, medium-, and low-resource directions.

¹<https://huggingface.co/datasets/NiuTrans/LMT-60-sft-data>

²<https://opus.nlpl.eu/ALT/en&my/v20191206/ALT>

For evaluation, we use FLORES+³ (NLLB Team et al., 2024; Goyal et al., 2022) as the test benchmark. FLORES+ extends FLORES-200 to a broader multilingual setting and is widely adopted for standardized multilingual translation evaluation. The dataset statistics are summarized in Table 1.

	en-vi	en-zh	en-ru	en-cs	en-de
Train	18K	25K	31K	27K	32K
Test	1K	1K	1K	1K	1K

Table 1: Dataset statistics for bilingual En→X parallel corpora; X→En splits are symmetric and omitted.

4.2 Model Settings

We evaluate MeRP on the fully frozen Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B⁴ (Yang et al., 2025) backbones, updating only the MeRP module. Unless otherwise noted, the main text reports the Gemma-2-2B-It setting, while the remaining configurations are provided in Appendix A.

For Gemma-2-2B-It, the prompt length is 32, the learning rate is 6×10^{-4} , and the shared trunk dimension is 256. The routed expert module uses 10 primary experts and 2 overflow experts, with expert rank $r = 16$ and direction embedding dimension 64. The warm-start scaling coefficient is 1.1×10^{-3} . The contrastive loss weight is 0.20, and the anchor loss weight is linearly annealed from 0.30 to 0.05, with minimum lead 0.14, margin weight 0.25, and target probabilities 0.87 and 0.065 for primary and overflow experts, respectively. Training uses batch size 4, gradient accumulation 32, 3 epochs, and warmup ratio 0.08.

We report BLEU, spBLEU (Goyal et al., 2022), chrF++, and COMET-22. Table 2 reports BLEU and COMET-22, and the remaining metrics are provided in the Appendix B. To assess MeRP from the perspectives of parameter count, efficiency, and translation quality, we further compare it with several representative PEFT baselines.

- **LoRA** (Hu et al., 2022): Low-rank updates are applied to the Q and V projections in self-attention, and a shared LoRA configuration is used across all language pairs.

³https://huggingface.co/datasets/openlanguage_data/flores_plus

⁴<https://huggingface.co/Qwen/Qwen3-4B>

- **P-tuningV1** (Liu et al., 2024): A prompt-based tuning method that inserts trainable continuous prompt embeddings into the input layer, utilizing a prompt encoder (e.g., an LSTM or MLP) to model token dependency for downstream tasks while keeping the language model frozen.
- **Adapter** (Ding et al., 2023): Adapter modules are inserted into the backbone model, and only the parameters of the Adapter modules are updated, with all other parameters frozen.
- **Prefix-Tuning** (Li and Liang, 2021): Task-specific vectors are prepended to the keys (K) and values (V) of the multi-head attention layers at every transformer block, and only these prefix parameters are optimized while keeping the backbone model frozen.

4.3 Main Results

Table 2 summarizes the BLEU and COMET-22 evaluation results for Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B across ten translation directions. The experimental results demonstrate that the proposed MeRP exhibits highly competitive performance across base models of different architectures and scales, particularly offering significant advantages in parameter efficiency and hardware adaptability.

On the Gemma backbone, MeRP shows strong parameter efficiency. Compared with P-tuningV1, which also belongs to input-layer adaptation, MeRP achieves superior performance in most translation directions, while its trainable parameter count is only 21.90% of P-tuningV1. Compared with LoRA, MeRP maintains comparable or better translation quality while reducing computational overhead. Its average BLEU and COMET-22 scores are on par with LoRA. Regarding efficiency metrics, the training time of MeRP is reduced by 37.48% compared to LoRA, and its trainable parameters are only 57.18% of the latter, achieving a superior performance-efficiency trade-off.

On the LLaMA backbone, although the optimal results for different translation directions are distributed across various PEFT methods, MeRP shows competitive overall performance. Specifically, MeRP achieved the best results in 5 translation directions for both BLEU and COMET-22 metrics. In contrast to the performance fluctuations of other baseline methods on specific language pairs, MeRP performs more balanced in mul-

Methods	BLEU/COMET-22										
	de-en	cs-en	zh-en	ru-en	vi-en	en-de	en-cs	en-zh	en-ru	en-vi	Avg.
<i>Gemma-2-2B-It</i>											
Base	36.58/88.12	34.36/87.19	20.73/85.01	30.80/85.56	10.91/59.02	20.18/81.09	12.37/80.87	10.26/52.05	10.40/74.39	11.22/52.49	19.78/74.58
LoRA	42.85/89.08	38.85/87.76	27.64/86.62	35.50/86.42	36.27/87.10	33.37/86.80	22.31/87.46	38.88/86.91	26.15/87.67	36.35/87.07	33.82/87.29
Prefix	42.58/89.09	37.84/87.70	27.78/86.50	34.47/86.29	35.83/86.87	30.23/83.31	21.75/87.27	37.99/86.09	25.88/86.17	35.07/85.16	32.94/86.45
Adapter	42.73/88.99	38.09/87.72	27.56/86.52	35.45/86.40	35.99/87.00	33.04/86.51	21.48/87.49	38.25/86.67	26.14/87.48	36.37/87.10	33.51/87.19
P-tuning	41.59/88.99	37.22/87.70	26.84/86.55	34.45/86.20	34.41/86.76	31.62/86.03	21.04/87.25	37.56/86.00	25.41/86.81	34.19/86.18	32.43/86.84
MeRP	42.92/89.12	38.97/87.90	28.58/86.76	35.43/86.38	36.43/87.15	33.24/86.75	22.21/87.61	38.99/87.20	26.13/87.60	36.35/87.08	33.93/87.36
<i>LLaMA-3.2-3B-Instruct</i>											
Base	30.08/80.54	24.71/79.49	9.97/73.93	23.00/77.07	6.59/55.87	10.02/72.48	8.53/72.67	11.21/70.38	6.16/63.93	11.42/50.21	14.17/69.66
LoRA	41.66/88.81	36.93/87.57	26.90/86.40	33.90/86.06	34.40/86.71	32.18/86.28	19.75/84.12	35.15/85.47	22.46/83.24	36.77/87.22	32.01/86.19
Prefix	40.82/88.55	36.82/87.48	25.72/86.34	33.78/86.05	33.81/86.49	32.21/86.30	20.98/85.24	35.60/85.56	22.84/83.12	37.07/87.40	31.96/86.25
Adapter	42.35/88.88	36.70/87.48	26.41/86.43	33.55/86.19	34.71/86.77	33.08/86.27	20.81/84.39	35.11/85.61	22.16/81.82	37.02/87.18	32.12/86.10
P-tuning	41.40/88.78	36.35/87.26	26.39/86.24	33.48/85.86	33.83/86.74	32.08/86.25	18.29/81.24	34.95/85.29	20.60/80.37	36.06/87.00	31.34/85.50
MeRP	42.00/88.89	36.85/87.47	26.96/86.36	33.88/86.12	35.01/86.89	32.68/86.52	20.58/84.46	35.73/85.77	22.88/83.65	36.76/87.31	32.33/86.34
<i>Qwen3-4B</i>											
Base	39.99/88.83	34.90/87.41	26.61/87.08	32.45/86.19	10.87/54.68	29.32/85.47	18.16/83.37	43.07/88.57	24.55/87.24	15.31/48.67	27.52/79.75
LoRA	42.42/89.14	37.11/87.74	29.03/87.14	33.64/86.30	35.72/87.15	31.76/86.24	20.68/86.09	43.08/88.30	25.30/88.07	38.11/87.89	33.68/87.41
Prefix	41.93/89.04	37.07/87.78	28.61/87.00	33.49/86.17	35.33/86.98	31.62/85.98	20.39/85.49	41.91/87.18	25.24/87.09	37.92/87.92	33.35/87.06
Adapter	41.85/89.01	37.02/87.65	29.03/87.01	34.18/86.41	35.20/87.05	30.90/85.61	19.58/85.66	42.70/88.30	25.00/87.65	37.87/87.90	33.33/87.23
P-tuning	41.43/88.99	36.49/87.58	29.02/87.19	33.52/86.24	34.75/86.95	30.73/86.13	18.33/85.07	42.53/88.38	24.89/88.07	37.46/88.17	32.92/87.28
MeRP	42.34/89.12	37.34/87.76	29.49/87.25	34.57/86.48	36.45/87.25	31.73/86.23	20.62/86.00	43.23/88.72	25.48/88.20	38.18/88.13	33.94/87.51

Table 2: Performance comparison of Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B on FLORES+ (BLEU/COMET-22) across ten translation directions; bold marks best per backbone.

tiling translation tasks by virtue of its direction-adaptive expert mechanism.

On the larger Qwen3-4B model, the memory-efficiency advantages of MeRP become even more evident. Under the same hardware constraints (A100 40GB), we adjusted batch size for each method to ensure training feasibility. LoRA and Adapter required a smaller batch size of 2 with 64 gradient accumulation steps due to memory constraints, while MeRP maintained batch size 4 without OOM. In contrast, MeRP completed training stably with its lightweight design and achieved the highest BLEU scores in 7 directions.

Methods	cs-en	de-en	zh-en	ru-en	vi-en
<i>BLEU</i>					
Base model	34.36	36.58	20.73	30.80	10.91
MeRP	38.97	42.92	28.58	35.43	36.43
w/o routed experts	38.57	43.02	27.54	34.83	36.13
w/o soft anchoring	38.34	42.65	27.55	34.85	36.37
w/o contrastive routing	38.39	42.60	27.92	34.76	36.46
<i>COMET-22</i>					
Base model	87.19	88.12	85.01	85.56	59.02
MeRP	87.90	89.12	86.76	86.38	87.15
w/o routed experts	87.81	89.06	86.52	86.33	87.03
w/o soft anchoring	87.72	89.06	86.53	86.24	87.08
w/o contrastive routing	87.71	89.04	86.54	86.28	87.02

Table 3: Ablation results of Gemma-2-2B-It on the X→EN task, comparing the full MeRP model with variants lacking key components.

Overall, the experimental results verify the efficiency and practical value of MeRP in multilingual adaptation of frozen LLM backbones. Especially in memory-constrained environments, MeRP provides a parameter-efficient adaptation scheme with favorable hardware feasibility.

4.4 Ablation Study

To evaluate the contributions of the various components of MeRP, we established three sets of ablation experiments, respectively removing the direction-adaptive experts, the soft anchoring mechanism, and the contrastive routing regularization. Table 3 reports the ablation results for Gemma-2-2B-It on the X→EN subset; comprehensive ablation data are provided in the Appendix C.

Overall, removing any major component generally degrades performance, indicating that the shared trunk, routed experts, and routing regularization work synergistically. At the same time, the contribution of each component varies across language pairs. For DE-EN, removing direction-adaptive experts yields slightly better results, suggesting that shared translation priors may already be sufficient for closely related languages.

Nevertheless, considering all evaluated directions collectively, the full MeRP model remains the most consistently competitive configuration, achieving the optimal overall balance between universal knowledge extraction and direction-specific

specialization modeling.

5 Analysis

5.1 Efficiency and Performance Analysis

Figure 3 presents a comparative analysis of trainable parameter size and translation performance (COMET-22 score) for different PEFT methods on Qwen3-4B. Among all evaluated methods, our proposed MeRP achieves the best overall performance; it not only delivers the highest COMET-22 score (87.51) but also keeps the trainable parameters to a minimum (only 1.96M). Specifically, compared to the popular baseline LoRA, MeRP significantly reduces the trainable parameters by approximately 66.7% while improving the translation quality. Overall, MeRP can effectively unlock the foundational model’s translation potential with minimal computational and parameter costs, achieving a dual breakthrough in both fine-tuning efficiency and model performance.

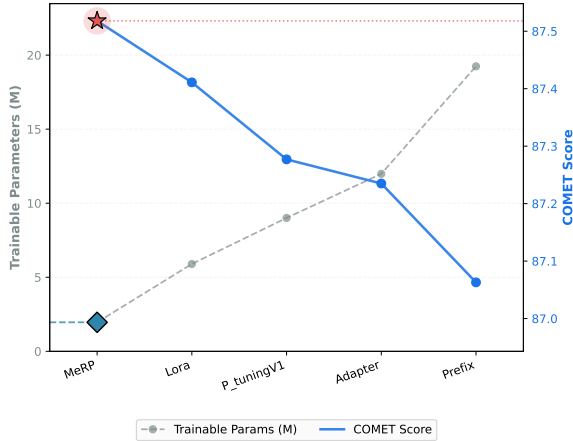


Figure 3: Parameter efficiency vs. COMET-22 performance of PEFT methods on Qwen3-4B. MeRP (star/diamond markers) achieves the best trade-off between translation quality and parameter overhead.

5.2 Directional Routing Analysis

Figure 4 visualizes the centered cosine similarity of routing patterns across different translation directions. The diagonal values are all 1, while the low off-diagonal similarities indicate that different translation directions tend to form differentiated expert combinations, supporting the learning of direction-specific expert preferences in the routing space. Soft Anchoring provides flexible expert preferences for different translation directions, guiding the model to favor the corresponding primary experts during the early stages of training,

while the final expert weights are adaptively determined by the model based on the input and training process. The contrastive routing regularization further promotes differentiated routing patterns across directions. Combined with the ablation results in Tables 3 and 8, where removing the routed experts, soft anchoring, or contrastive routing regularization leads to a decrease in average performance, these results further demonstrate the positive contribution of direction-aware routing to adaptation across different translation directions.

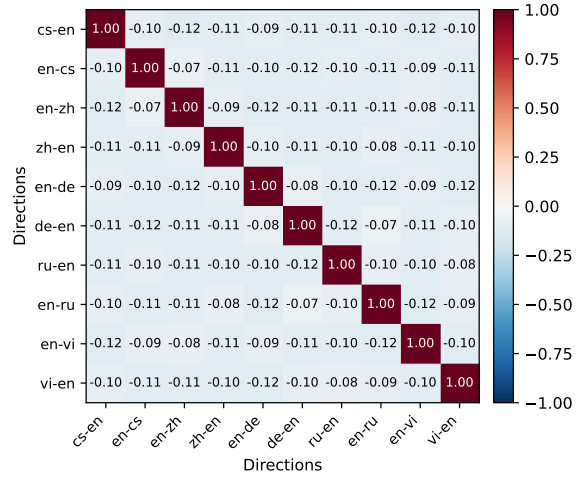


Figure 4: Routing analysis of MeRP on a single backbone model. Centered cosine similarity of routing patterns, where low off-diagonal similarity indicates directional separation in routing space.

5.3 Cross-Directional Interference in Large-Scale Language Models

To examine whether cross-directional negative transfer persists in relatively large-scale modern LLMs, we further evaluate Gemma-3-27B-It, a large open-source LLM that is deployable under our available computational resources. We compare single-direction training using only English→Vietnamese data with mixed-direction training using English→Vietnamese, Vietnamese→English, English→Burmese, and Burmese→English data. Both models are evaluated on the same English→Vietnamese test set. As shown in Table 4, mixed-direction training consistently degrades performance, with BLEU decreasing from 45.39 to 43.22 and COMET-22 from 89.53 to 88.41. These results indicate that cross-directional negative transfer can still occur when different translation directions share adaptation parameters, even in relatively large-scale modern

LLMs. Training hyperparameters are provided in Appendix G.

Methods	en-vi
<i>BLEU/chrF+/COMET-22</i>	
Single-direction	45.39/61.28/89.53
Mixed-direction	43.22/60.53/88.41

Table 4: Performance of LoRA on the EN→VI task with single-direction and mixed-direction training using Gemma-3-27B-It.

5.4 Many-to-One and One-to-Many Translation

We further analyze the performance of MeRP under the Many-to-One (M2O) and One-to-Many (O2M) training settings. The M2O setting includes only translation directions with English as the target language, while the O2M setting includes only directions with English as the source language. Experiments are conducted on the Gemma-2-2B-It and Qwen3-4B backbones using the same training configurations as the main experiments. As shown in Table 5, under the M2O setting, MeRP achieves the highest average scores on both backbones. Under the O2M setting, MeRP achieves the best average performance on Qwen3-4B. Overall, the average performance of O2M is lower than that of M2O, indicating that, under the current setting, translating from a single source language into multiple heterogeneous target languages presents greater translation difficulty than translating from multiple source languages into a single target language.

Methods	M2O	O2M
<i>Gemma-2-2B-It</i>		
LoRA	36.33/87.41	31.12/ 87.24
Prefix	35.82/87.32	31.23 /87.22
Adapter	36.10/87.34	31.13/87.22
P-tuningV1	35.52/87.23	30.04/86.78
MeRP	36.44/87.42	31.03/87.17
<i>Qwen3-4B</i>		
LoRA	36.03/87.59	31.64/87.33
Prefix	36.03/87.58	31.60/87.16
Adapter	35.80/87.50	31.47/87.22
P-tuningV1	34.83/87.41	30.89/87.01
MeRP	36.23/87.64	31.75/87.36

Table 5: Performance comparison of PEFT methods under Many-to-One (M2O) and One-to-Many (O2M) settings on FLORES+ (BLEU/COMET-22).

6 Conclusion

This paper presents MeRP (Multi-Expert Routing Prompt), a multi-expert routing prompt encoding strategy for PEFT in multilingual machine translation. MeRP encodes source-side representations through a shared cross-attention trunk and dynamically computes soft routing distributions over low-rank experts according to the translation direction. This design enables MeRP to extract shared translation priors while modeling direction-specific residual information. Experiments on Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B show that MeRP achieves competitive or even superior multilingual translation performance compared with representative PEFT methods, while requiring substantially fewer trainable parameters. Ablation studies demonstrate the effectiveness of the routed experts, soft anchoring, and contrastive routing regularization. The results show that these components contribute to direction-specific adaptation and improved multilingual translation performance.

Limitations

Our experiments have not yet evaluated the applicability of MeRP to translation directions centered on languages other than English or to the latest generation of large-scale models. In addition, its performance gains may be limited for languages with relatively scarce pre-training resources. Future work will further evaluate the robustness and scalability of MeRP across broader multilingual settings and newer, larger-scale models.

Ethics Statement

This research follows an ethical framework and does not involve ethical issues. The data constructed for this research is derived from open-source data, and the large language models employed in this study follow their declared licenses.

Acknowledgments

We thank all reviewers for their helpful comments. This work financially supported by the National Natural Science Foundation of China (62666040, 62366027, 62306129, 62266027, U23A2038, U24A20334), Yunnan Fundamental Research Projects (202601CI070148, 202401CF070121, 202401BC070021), Major Science and Technology Special Projects of Yunnan Province (202602AD080015, 202502AD080014,

202502AD080016, 202402AG050007, 202402AD080002), Kunming University of Science and Technology Construction Joint Project (2023 01BE070001-027), Major Science and Technology Achievement Transformation Project of Yunnan Provincial Department of Education (FWCY-ZD2025006), Intergovernmental International Science, Technology and Innovation Cooperation (2025YFE0121700)

References

- Duarte Alves, Nuno Guerreiro, João Alves, José Pomal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. [Multilingual machine translation with open large language models at practical scale: An empirical study](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico.
- Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.
- Tianyu Dong, Bo Li, Jinsong Liu, Shaolin Zhu, and Deyi Xiong. 2025. [MLAS-LoRA: Language-aware parameters detection and LoRA-based knowledge transfer for multilingual machine translation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15645–15660, Vienna, Austria.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. 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.
- Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models. *arXiv preprint arXiv:2401.05861*.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2024. Gpt understands, too. *AI open*, 5:208–215.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).
- Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863.
- Tong Su, Xin Peng, Sarubi Thillainathan, David Guzmán, Surangika Ranathunga, and En-Shiun Lee. 2024. Unlocking parameter-efficient fine-tuning for low-resource language translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4217–4225.
- Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into chatgpt’s customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204.

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

Yuxuan Ouyang Murun Yang Dingyang Lin Kaiyan Chang Tong Zheng Bei Li Peinan Feng Quan Du Tong Xiao Jingbo Zhu Yingfeng Luo, Ziqiang Xu. 2025. [Beyond english: Toward inclusive and scalable multilingual machine translation with llms](#). *Preprint*, arXiv:2511.07003.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

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*.

Tong Zheng, Yan Wen, Huiwen Bao, Junfeng Guo, and Heng Huang. 2025. Asymmetric conflict and synergy in post-training for llm-based multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18362–18383.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024a. [LANDeRMT: Detecting and routing language-aware neurons for selectively finetuning LLMs to machine translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148, Bangkok, Thailand.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. [Multilingual machine translation with large language models: Empirical results and analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2812–2823.

A Main Experimental Configurations

Table 6 summarizes the principal experimental configurations used for the three backbone models. Across all settings, MeRP keeps the same core architecture, including prompt length, shared trunk width, expert count, expert rank, and direction embedding size, which helps isolate the effect of the backbone model itself. The main differences lie in optimization-sensitive hyperparameters, especially the prompt encoder learning rate, initialization scale of the up-projection, and the strength of

routing regularization. In particular, the LLaMA-3.2-3B-Instruct configuration uses a larger learning rate and weaker anchoring targets, while the Gemma-2-2B-It and Qwen3-4B settings adopt relatively stronger anchoring to encourage clearer expert specialization.

B Additional Translation Metrics

Table 7 complements the main-text BLEU and COMET-22 results with spBLEU and chrF++ scores. The overall trends remain consistent with the main results, and MeRP continues to achieve better multilingual translation performance under a tight parameter budget. On Gemma-2-2B-It, MeRP achieves the best average chrF++ and remains essentially tied with LoRA in average spBLEU, indicating that the proposed routing design does not sacrifice surface-form fidelity. On LLaMA-3.2-3B-Instruct, MeRP attains the best average spBLEU and remains close to the strongest chrF++ baseline, suggesting balanced gains across token-level and character-level evaluation. On Qwen3-4B, MeRP yields the best average chrF++ and highly competitive average spBLEU, further confirming that its advantages extend beyond BLEU and COMET-22 to complementary translation metrics.

C Ablation Results

Table 8 extends the ablation study from Gemma-2-2B-It to all three backbone models, including LLaMA-3.2-3B-Instruct and Qwen3-4B. Across all settings, the full MeRP model consistently achieves the best average BLEU and COMET-22 scores, indicating that routed experts, soft anchoring, and contrastive routing regularization are complementary rather than redundant components. Among the ablations, removing routed experts leads to the most pronounced degradation in average BLEU, suggesting that direction-specific residual modeling is crucial for translation quality, while removing soft anchoring or contrastive routing results in more moderate but consistent performance drops across all backbones, reflecting their role in stabilizing expert routing and improving cross-direction consistency. Overall, these trends across both metrics and model scales support the effectiveness of MeRP in jointly leveraging shared translation priors and structured direction-aware specialization.

Hyperparameter	Gemma-2-2B-It	LLaMA-3.2-3B-Instruct	Qwen3-4B
Prompt length P	32	32	32
Prompt encoder LR	6×10^{-4}	2×10^{-3}	8×10^{-4}
Shared trunk dim d_t	256	256	256
Experts (K)	12 (10+2 overflow)	12 (10+2 overflow)	12 (10+2 overflow)
Expert rank r	16	16	16
Direction dim d_{dir}	64	64	64
Up-proj init std	1.1×10^{-3}	4×10^{-4}	1.0×10^{-3}
Contrastive weight	0.20	0.14	0.16
Anchor weight schedule	$0.30 \rightarrow 0.05$	$0.12 \rightarrow 0.01$	$0.20 \rightarrow 0.01$
Anchor target (main / overflow)	0.87 / 0.065	0.77 / 0.10	0.85 / 0.05
Anchor margin / weight	0.14 / 0.25	0.09 / 0.14	0.12 / 0.21
Epochs	3	3	3
Warmup ratio	0.08	0.10	0.08
Batch size / grad. accum.	4 / 32	4 / 32	4 / 32

Table 6: Main training configurations used for the three backbone models in the principal MeRP experiments. The table summarizes the key architecture and optimization hyperparameters adopted for Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B. All settings use frozen backbones, ten translation directions, BF16 training, and 12 routed experts with two overflow experts.

Methods	spBLEU/chrF++										
	de-en	cs-en	zh-en	ru-en	vi-en	en-de	en-cs	en-zh	en-ru	en-vi	Avg.
<i>Gemma-2-2B-It</i>											
Base	40.59/63.85	37.97/60.95	23.62/52.19	34.18/58.54	13.10/45.07	27.72/53.67	19.16/42.11	6.95/11.81	16.66/41.46	11.22/38.00	23.12/46.77
LoRA	47.24/65.98	42.54/62.47	32.47/ 54.81	39.52/60.29	39.64/59.38	39.75/58.85	29.19/47.90	32.04/28.50	33.77/51.09	36.35/54.37	37.25/54.36
Prefix	47.07/65.94	41.90/62.21	31.90/54.52	38.55/59.81	39.12/58.76	35.91/54.62	28.68/47.40	31.36/28.55	33.07/49.93	35.07/52.24	36.26/53.40
Adapter	47.01/65.75	41.91/62.04	31.80/54.43	39.48/60.10	39.33/59.06	39.43/58.24	29.06/47.50	31.56/27.90	33.67/50.92	36.37/54.41	36.96/54.04
P-tuning	46.10/65.72	41.30/61.94	31.03/54.59	38.40/59.80	37.92/57.91	38.15/57.39	27.64/46.57	30.42/26.93	32.78/49.85	34.19/52.01	35.79/53.27
MeRP	47.24/65.86	42.66/62.63	32.56/54.80	39.31/60.22	39.73/59.78	39.45/58.69	29.10/47.51	32.43/29.06	33.66/50.84	36.35/ 54.61	37.25/54.40
<i>LLaMA-3.2-3B-Instruct</i>											
Base	32.22/54.56	28.93/50.44	13.69/28.83	26.62/47.08	7.92/34.04	14.14/38.89	13.45/33.04	7.13/9.12	10.04/25.07	11.42/37.18	16.56/35.82
LoRA	46.19/65.05	40.74/61.17	31.00/ 53.83	37.89/59.12	37.96/58.10	38.31/57.78	27.28/45.58	28.27/26.01	29.34/46.56	36.77/54.48	35.38/52.77
Prefix	45.06/64.56	41.15/ 61.70	29.70/53.49	37.74/ 59.45	37.47/57.54	39.08/58.27	28.50/46.45	29.06/ 27.69	30.08/47.12	37.07/54.93	35.49/ 53.12
Adapter	46.77/65.33	41.10/61.48	30.55/53.52	37.99/59.37	38.36/58.51	38.51/57.76	27.92/46.09	28.29/26.12	28.66/45.31	37.02/54.72	35.52/52.82
P-tuning	45.79/65.17	41.21/61.69	30.61/53.77	37.68/59.22	37.69/57.99	38.35/57.91	25.49/43.43	27.61/24.74	27.18/43.53	36.26/54.10	34.79/52.16
MeRP	46.42/65.15	40.68/61.24	31.30/53.81	37.58/59.43	38.41/58.74	38.85/57.88	27.21/46.02	29.90/26.79	30.10/47.25	36.76/54.79	35.72/53.11
<i>Qwen3-4B</i>											
model	44.26/64.89	38.86/60.94	30.94/55.37	36.20/58.97	12.77/45.63	35.48/55.90	24.65/44.37	34.69/27.86	31.82/49.81	16.31/44.06	30.60/50.78
LoRA	46.95/65.73	41.03/61.49	33.61/55.78	37.60/59.73	39.18/59.16	37.66/57.41	27.40/46.38	36.29/33.04	32.76/ 50.54	38.11/55.90	37.06/54.52
Prefix	46.46/65.46	40.93/61.59	32.63/55.20	37.33/59.65	38.45/58.53	37.30/57.14	27.03/45.92	35.24/30.21	32.59/49.69	37.92/55.59	36.59/53.90
Adapter	46.40/65.39	41.15/61.65	33.83/56.17	38.55/59.91	38.51/58.80	36.73/56.53	25.54/45.71	36.01/31.49	32.40/50.27	37.87/55.66	36.70/54.16
P-tuning	45.91/65.47	40.55/61.47	33.44/55.80	37.55/59.50	38.12/58.57	36.79/56.83	24.49/44.66	35.62/29.32	32.40/50.20	37.46/55.43	36.23/53.73
MeRP	46.68/65.51	41.28/61.80	33.79/55.96	38.68/59.94	39.48/59.27	37.45/57.50	27.20/45.90	36.62/33.97	32.80/50.41	38.23/56.05	37.22/54.63

Table 7: Performance comparison of Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B on the FLORES+ test set using spBLEU and chrF++. The table reports ten translation directions and the corresponding averages, with bold indicating the best result for each backbone model in each direction or average column.

D Additional Routing Analysis

Figure 5 extends the routing analysis in the main text from Gemma-2-2B-It to LLaMA-3.2-3B-Instruct. The routing heatmap again shows clear direction-dependent expert preferences rather than uniform expert usage, indicating that MeRP preserves backbone-agnostic specialization behavior. The similarity matrix exhibits relatively low off-diagonal values, suggesting that different translation directions remain well separated in routing space even when the backbone changes. At the same time, the anchor-alignment panel shows that the main expert consistently retains dominant probability mass while overflow experts absorb a limited amount of shared capacity, which is consistent

with the soft-anchoring design. The entropy panel further indicates sparse expert allocation, supporting the conclusion that MeRP reduces multilingual interference on LLaMA by combining stable routing preferences with flexible residual sharing.

Figure 6 presents the corresponding routing analysis for Qwen3-4B. Despite the larger backbone scale, MeRP still learns structured and direction-sensitive expert usage. The routing heatmap shows that multiple translation directions develop stable but distinct expert preferences, while the similarity panel indicates limited overlap among routing patterns for different directions. This behavior is important for Qwen because the main experiments show that MeRP remains especially attractive under

(a)											
Methods	de-en	cs-en	zh-en	ru-en	vi-en	en-de	en-cs	en-zh	en-ru	en-vi	Avg.
<i>Gemma-2-2B-It</i>											
Base model	36.58	34.36	20.73	30.80	10.91	20.18	12.37	10.26	10.40	11.22	19.78
MeRP	42.92	38.97	28.58	35.43	36.43	33.24	22.21	38.99	26.13	36.35	33.93
w/o routed experts	43.02	38.57	27.54	34.83	36.13	32.47	21.03	38.33	26.24	35.28	33.34
w/o soft anchoring	42.65	38.34	27.55	34.85	36.37	32.43	21.51	38.53	25.98	35.79	33.40
w/o contrastive routing	42.60	38.39	27.92	34.76	36.46	32.55	21.06	38.49	25.71	35.70	33.36
<i>LLaMA-3.2-3B-Instruct</i>											
Base model	30.08	24.71	9.97	23.00	6.59	10.02	8.53	11.21	6.16	11.42	14.17
MeRP	42.00	36.85	26.96	33.88	35.01	32.68	20.58	35.73	22.88	36.76	32.33
w/o routed experts	41.85	36.51	26.82	33.46	34.92	32.04	20.05	34.34	22.67	35.74	31.84
w/o soft anchoring	42.09	36.46	26.90	33.92	34.80	32.38	20.26	35.50	22.47	36.43	32.12
w/o contrastive routing	41.96	36.67	26.44	33.66	34.84	32.61	20.50	35.73	22.51	36.37	32.13
<i>Qwen3-4B</i>											
Base model	39.99	34.90	26.61	32.45	10.87	29.32	18.16	43.07	24.55	15.31	27.52
MeRP	42.34	37.34	29.49	34.57	36.45	31.73	20.62	43.23	25.48	38.18	33.94
w/o routed experts	42.02	37.51	28.58	34.41	36.05	31.32	19.34	42.95	25.21	37.54	33.49
w/o soft anchoring	42.45	37.44	28.92	34.23	36.33	31.75	20.10	43.08	25.40	38.09	33.78
w/o contrastive routing	42.38	37.29	29.19	34.54	36.30	31.59	20.12	43.13	25.53	38.20	33.83
(b)											
Methods	de-en	cs-en	zh-en	ru-en	vi-en	en-de	en-cs	en-zh	en-ru	en-vi	Avg.
<i>Gemma-2-2B-It</i>											
Base model	88.12	87.19	85.01	85.56	59.02	81.09	80.87	52.05	74.39	52.49	74.58
MeRP	89.12	87.90	86.76	86.38	87.15	86.75	87.61	87.20	87.60	87.08	87.36
w/o routed experts	89.06	87.81	86.52	86.33	87.03	86.51	87.15	86.93	87.47	86.91	87.17
w/o soft anchoring	89.06	87.72	86.53	86.24	87.08	86.44	87.02	86.74	87.49	87.00	87.13
w/o contrastive routing	89.04	87.71	86.54	86.28	87.02	86.52	87.25	86.76	87.55	86.97	87.16
<i>LLaMA-3.2-3B-Instruct</i>											
Base model	80.54	79.49	73.93	77.07	55.87	72.48	72.67	70.38	63.93	50.21	69.66
MeRP	88.89	87.47	86.36	86.12	86.89	86.52	84.46	85.77	83.65	87.31	86.34
w/o routed experts	88.77	87.41	86.31	85.90	86.77	86.20	83.22	84.69	83.45	86.57	85.93
w/o soft anchoring	88.82	87.40	86.31	86.02	86.70	86.33	84.10	85.78	83.58	87.33	86.24
w/o contrastive routing	88.89	87.39	86.19	85.93	86.74	86.41	84.23	85.70	83.56	87.24	86.23
<i>Qwen3-4B</i>											
Base model	88.83	87.41	87.08	86.19	54.68	85.47	83.37	88.57	87.24	48.67	79.75
MeRP	89.12	87.76	87.25	86.48	87.25	86.23	86.00	88.72	88.20	88.13	87.51
w/o routed experts	89.02	87.80	87.01	86.43	86.92	85.95	85.43	88.62	88.11	87.92	87.32
w/o soft anchoring	89.08	87.77	87.09	86.41	87.22	86.12	85.83	88.57	88.13	88.02	87.42
w/o contrastive routing	89.10	87.74	87.03	86.46	87.23	86.17	85.53	88.63	88.21	87.89	87.40

Table 8: Ablation study of MeRP on BLEU (a) and COMET-22 (b) scores across 10 translation directions for three base models, Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B. Best results are highlighted in bold.

stronger memory constraints; the appendix analysis suggests that this practical efficiency does not come at the cost of collapsed routing structure. In addition, the anchor-alignment and entropy panels show that routing remains concentrated and interpretable, which further supports the main-paper claim that MeRP maintains sparse specialization and practical trainability across different frozen LLM backbones.

E Trainable Parameter Comparison

As shown in Figure 7, we compare MeRP with four representative PEFT methods across three frozen backbone models in terms of trainable parameter

counts. MeRP consistently achieves the lowest number of trainable parameters among all settings, requiring only 1.83M, 2.22M, and 1.96M parameters on Gemma-2-2B-It, LLaMA-3.2-3B-Instruct, and Qwen3-4B, respectively. More specifically, MeRP reduces the average number of trainable parameters to 57.18% of LoRA and 21.90% of P-tuningV1, while accounting for less than 0.06% of the total parameters of the backbone models. This substantial efficiency gain remains stable across models of different scales and architectures, indicating that MeRP’s lightweight input-prefix design can significantly reduce fine-tuning overhead without requiring layer-wise modifications to the back-

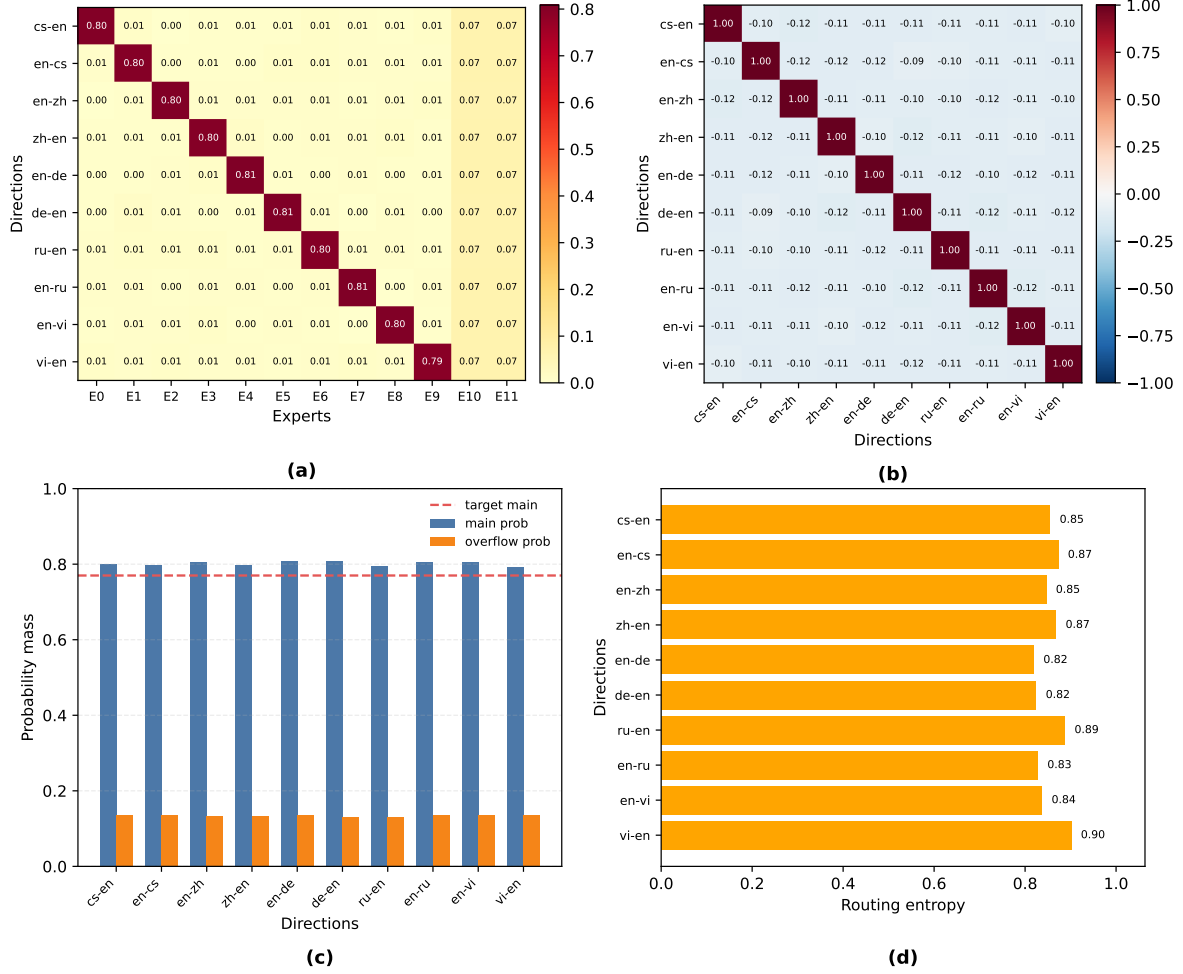


Figure 5: Routing analysis of MeRP on the LLaMA-3.2-3B-Instruct. (a) Routing heatmap across translation directions and experts, showing clear direction-aware expert preference. (b) Centered cosine similarity of routing patterns, where low off-diagonal similarity indicates directional separation in routing space. (c) Main-expert and overflow probability mass, showing that soft anchoring maintains a stable main-expert preference without enforcing a hard assignment. (d) Routing entropy across directions, indicating concentrated and sparse expert allocation.

bone.

Combined with the training speed improvements reported in this work, this lightweight property makes MeRP particularly suitable for memory-constrained multilingual machine translation scenarios, offering improved practicality and deployment efficiency.

F Training Times Comparison

The training time comparison is summarized in Figure 8. On Gemma-2-2B-It, MeRP requires only 5 h 27 min, corresponding to a 37.48% reduction compared with LoRA (8 h 43 min). On LLaMA-3.2-3B-Instruct, MeRP completes training in 4 h 50 min, yielding a 35.56% speedup over LoRA (7 h 30 min).

On Qwen3-4B, the efficiency advantage of

MeRP becomes more evident. Under the default batch size of 4×32 on a single 40 GB A100 GPU, LoRA, Prefix-Tuning, Adapter, and P-tuningV1 all encounter out-of-memory (OOM) issues, and therefore must be trained with a reduced batch size of 2×64 . In contrast, MeRP is the only method that remains stable under the 4×32 setting. Even under this more favorable configuration for the baselines, MeRP completes training in 7 h 13 min, significantly faster than LoRA (22 h), Prefix-Tuning (13 h 20 min), Adapter (14 h 58 min), and P-tuningV1 (13 h 52 min). Overall, MeRP reduces training time by 46.75% on average compared to LoRA across the three backbones. This gain is mainly due to its input-layer-only prefix design, which avoids the additional per-layer computation introduced by LoRA-style or adapter-based methods while reduc-

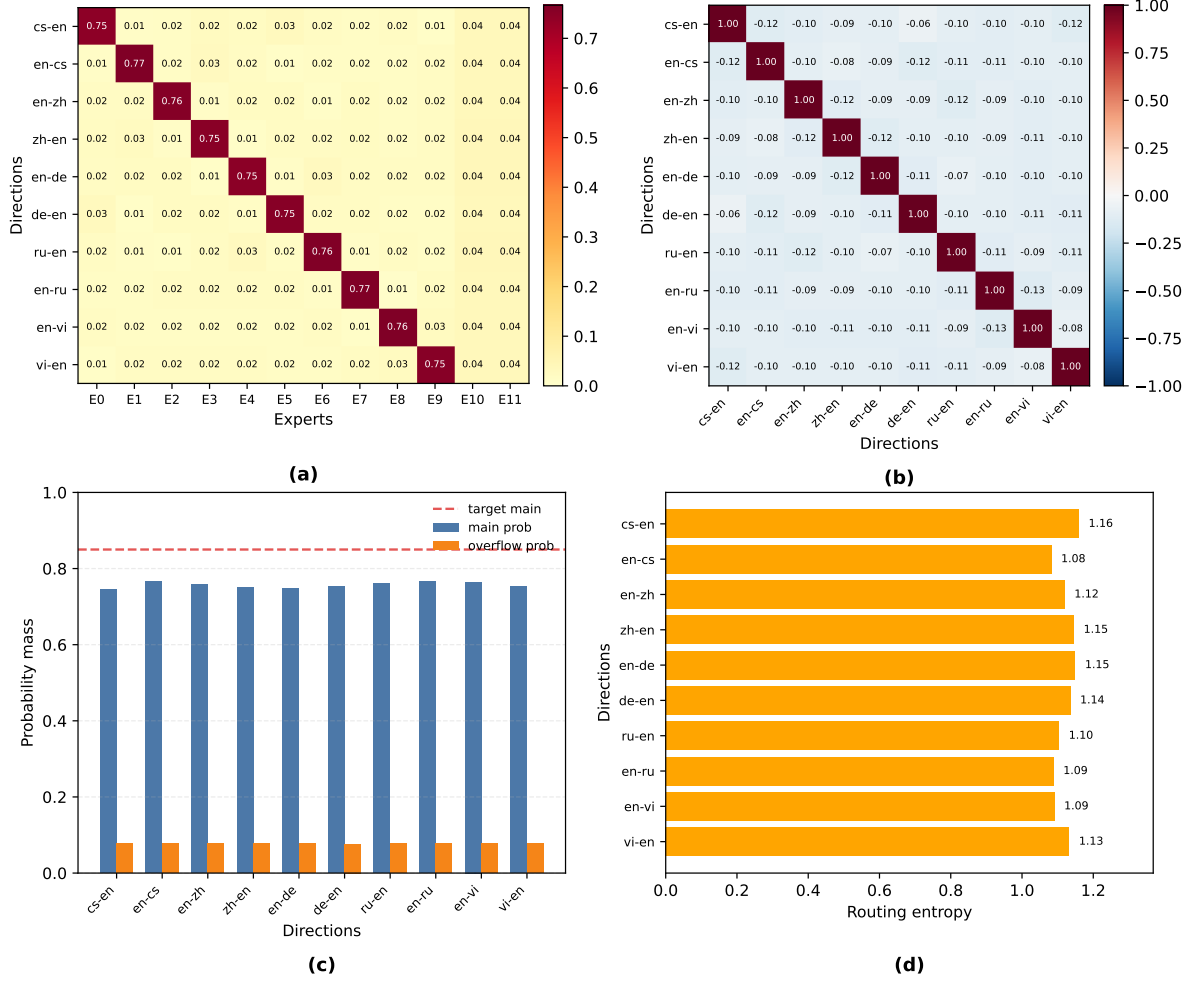


Figure 6: Routing analysis of MeRP on the Qwen3-4B. (a) Routing heatmap across translation directions and experts, showing clear direction-aware expert preference. (b) Centered cosine similarity of routing patterns, where low off-diagonal similarity indicates directional separation in routing space. (c) Main-expert and overflow probability mass, showing that soft anchoring maintains a stable main-expert preference without enforcing a hard assignment. (d) Routing entropy across directions, indicating concentrated and sparse expert allocation.

ing gradient update overhead. As a result, MeRP is particularly well suited for resource-constrained multilingual machine translation scenarios.

G Gemma-3-27B-It Experimental Details

For the Gemma-3-27B-It experiment, we use LoRA with a rank of 8 and apply it to the query and value projection matrices (q_{proj} and v_{proj}). Training is conducted on four GPUs with a batch size of 1 and 64 gradient accumulation steps. The model is trained for 2 epochs with a learning rate of 2×10^{-4} . Table 9 presents a representative English→Vietnamese translation example.

The single-direction model correctly translates “phosphorus” and “sulfur” into their corresponding Vietnamese terms, whereas the mixed-direction model directly retains the two English terms in the

Vietnamese output. This example illustrates the degradation in translation quality observed under mixed-direction training.

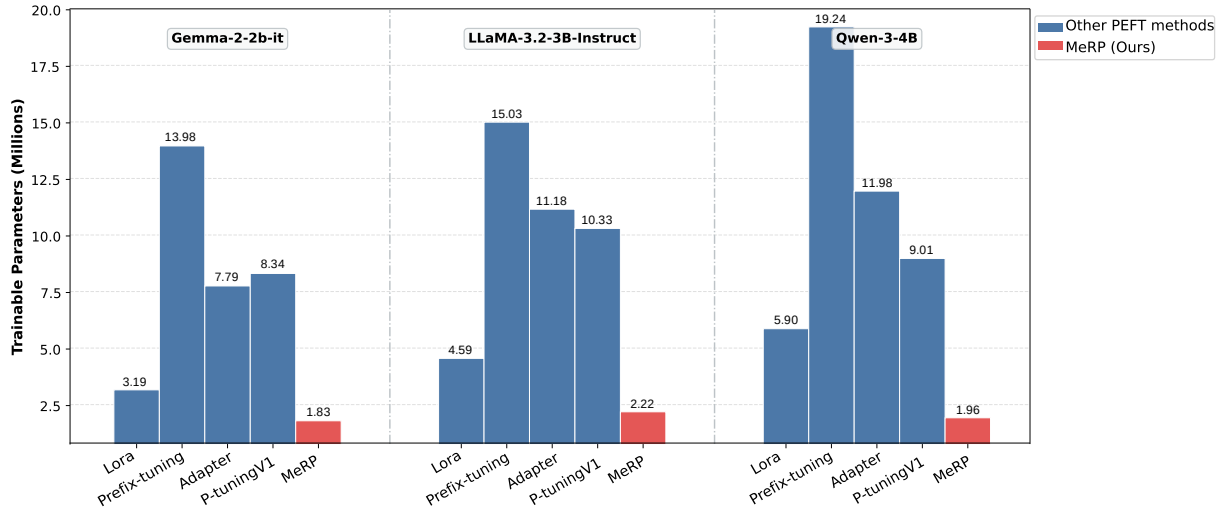


Figure 7: Trainable parameter counts (in millions) of different PEFT methods across three frozen LLM backbones. MeRP (red) consistently uses the fewest trainable parameters, demonstrating superior parameter efficiency.

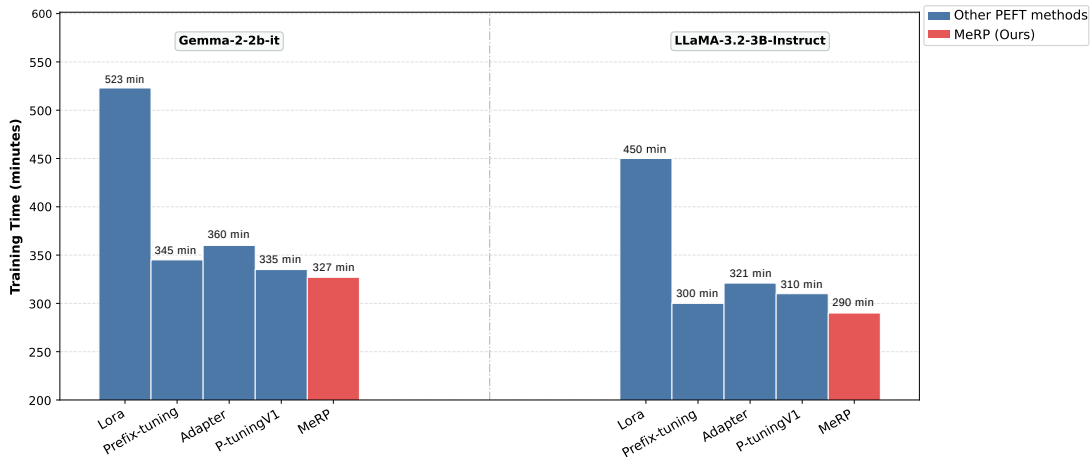


Figure 8: Comparison of training time between MeRP and four parameter-efficient fine-tuning methods on two frozen backbone models, Gemma-2-2B-It and LLaMA-3.2-3B-Instruct. Due to out-of-memory (OOM) errors under the batch size setting of 4×32 , LoRA and other PEFT baselines cannot be evaluated on Qwen3-4B; therefore, their training times on this model are not reported. MeRP (red) consistently achieves the shortest training time among all methods.

	Single-direction	Mixed-direction
Src	They did two experiments marking either the DNA in the phage with a radioactive phosphorus or the protein of the phage with radioactive sulfur.	
Translation	H đã thực hiện hai thí nghiệm đánh dấu học DNA trong phage vì pht pho phóng x học protein ca phage vì lu huỳnh phóng x.	H đã thực hiện hai thí nghiệm đánh dấu học DNA trong phage vì mt phosphorus phóng x học protein ca phage vì sulfur phóng x.
BLEU	38.77	27.13
COMET-22	80.60	73.82

Table 9: A representative English-to-Vietnamese translation example from Gemma-3-27B-It under single-direction and mixed-direction training.