DiET: Direction-specific Encoder Tuning for Large Language Models in Multilingual Machine Translation

Yuxin Gao[1], Zhenhan Wang[2]*, Fengzhao Fan[2]
[1]School of Software, Northeastern University, Shenyang, China
[2]Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
gaoyx2@mails.neu.edu.cn, 18640248002@163.com, fengzhao_fan@163.com

Abstract

The rapid development of large language models (LLMs) has reshaped machine translation (MT), enabling a single model to support multiple translation directions. However, LLM-based multilingual machine translation (MMT) still suffers from cross-lingual parameter interference when multiple directions share the same adaptation parameters. To address this issue, we propose DiET, a direction-specific encoder tuning method for MMT with LLMs. DiET deterministically selects a direction-specific encoder based on the source and target languages, which re-encodes the source text into input-dependent continuous prefix representations. These representations are concatenated with the translation instruction and source text and fed into a frozen LLM. During training, only the selected encoder and its projection modules are updated, providing separate adaptation spaces for different translation directions. We evaluate DiET on Qwen3-4B and Llama-3.2-3B-Instruct across 10 bidirectional translation directions between English and Burmese, Chinese, German, Russian, and Czech. On FLORES+ devtest, DiET improves over the frozen backbone by 1.81/0.48 and 3.87/4.60 BLEU/COMET, respectively, and outperforms strong parameter-efficient fine-tuning (PEFT) baselines on most directions. These results demonstrate that direction-specific encoders provide effective, input-dependent adaptation while keeping the LLM backbone frozen.

1 Introduction

LLMs have achieved state-of-the-art performance in MT in recent years, demonstrating remarkable cross-lingual generalization capabilities (Cui et al., 2025a; Zhu et al., 2024c; Lai et al., 2023). Benefiting from large-scale pre-training data, LLMs can effectively capture general linguistic patterns and exhibit strong multilingual capabilities. This inherent capability provides a solid foundation for MMT, making the adaptation of general-purpose LLMs an important research direction (Dong et al., 2025; Zhu et al., 2024a), while enabling multiple translation directions to be supported within a unified representation space.

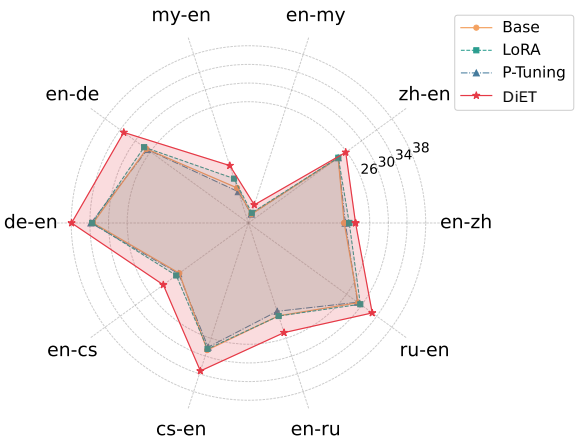


Figure 1: Multilingual translation performance across diverse language pairs based on Llama-3.2-3B-Instruct, evaluated with BLEU. Comparisons include Base (the frozen backbone model), LoRA, P-Tuning, and DiET, with DiET consistently yielding superior results across both high- and low-resource directions.

To effectively adapt these large-scale models to MMT tasks, PEFT has become a widely adopted paradigm. Methods such as LoRA (Hu et al., 2022a), Prefix-Tuning (Li and Liang, 2021a), and Adapter (Alves et al., 2023) freeze the backbone parameters while introducing only a small number of trainable parameters. This approach adapts LLMs to MT by optimizing the introduced parameters while keeping the backbone frozen.

However, applying standard PEFT methods to MMT introduces a fundamental challenge: cross-lingual parameter interference arising from linguistic heterogeneity. Existing PEFT methods typically share a set of global trainable parameters

*Corresponding author.

across languages. Due to substantial intrinsic differences among languages, forcing different translation directions, such as English→Chinese and English→German, to compete within a limited low-rank parameter space can lead to competition for model capacity. This phenomenon is often referred to as the curse of multilinguality, where conflicting gradient updates across languages allow high-resource languages to dominate representations needed by low-resource languages. As a result, models may suffer from capacity dilution, leading to degraded overall performance and limited ability to capture language-specific characteristics.

Previous studies have attempted to alleviate this issue by introducing language- or task-specific parameters. For example, language-specific Adapters add independent adaptation modules within the backbone network, while Prefix-Tuning learns continuous prefix parameters and injects them as additional key/value representations into the attention layers of the Transformer. However, the prefix parameters learned by classical Prefix-Tuning are typically independent of the specific input content, making it difficult to dynamically construct adaptation representations according to the source sentence. In contrast, DiET does not modify the internal Transformer architecture of the LLM. Instead, it introduces direction-specific encoders outside the backbone model to dynamically generate input-dependent continuous prefix representations based on the current source sentence.

We extensively evaluate DiET on multilingual benchmark datasets covering high-, medium-, and low-resource languages. The experimental results show that DiET significantly outperforms standard PEFT baselines (Figure 1) and effectively alleviates cross-lingual parameter interference. Further analysis demonstrates that explicitly modeling language-aware representations during adaptation enables the model to better handle linguistic diversity, providing an effective approach for adapting LLMs to multilingual scenarios.

Our contributions are summarized as follows:

- We propose DiET, a direction-specific encoder tuning method with a frozen backbone. DiET deterministically selects an encoder based on the translation direction, re-encodes the source sentence representation, and injects the resulting continuous prefix representations into the backbone input.
- We introduce a direction-specific representation construction and fusion approach. The approach first obtains source-sentence representations through the shared input embedding layer of the LLM, then employs a direction-specific encoder to generate input-dependent continuous prefixes, which are concatenated with the translation instruction and source-text embeddings for multilingual translation.
- We demonstrate the flexibility of direction-specific encoders in DiET. Different translation directions can employ encoders with different depths, enabling tailored modeling according to the characteristics of different language pairs and further improving multilingual translation performance.

2 Related Work

2.1 Multilingual Machine Translation with Large Language Models

In recent years, LLMs have demonstrated strong multilingual translation capabilities. Zhu et al. (2024d) conducted a large-scale evaluation of eight mainstream LLMs, including ChatGPT and GPT-4 (Achiam et al., 2023), covering 102 languages and 606 translation directions, and compared them with state-of-the-art translation systems such as No Language Left Behind (NLLB) (Costa-Jussà et al., 2022). Zhu et al. (2024b) proposed LANDeRMT, which detects language-specific and language-general neurons and selectively updates them to alleviate parameter interference. Lin et al. (2022) found that the 7.5B-parameter XGLM model, pretrained on 32 training examples, outperformed GPT-3 on 171 translation directions in FLORES-101 (Goyal et al., 2022a). Gao et al. (2024) proposed XConST, a regularization method that extends CrossConST (Gao et al., 2023) to better model cross-lingual differences, and investigated the impact of different prompting strategies on zero-shot translation performance. Cui et al. (2025b) further evaluated the multilingual translation capabilities of practical open-source LLMs with fewer than 10B parameters and proposed a Parallel-First Monolingual-Second (PFMS) data-mixing strategy.

Beyond directly using LLMs for translation, recent studies have also explored hybrid frameworks that combine LLMs with conventional neural machine translation (NMT) systems. In such frameworks, LLMs serve as collaborative decoders to

further refine initial translations generated by NMT systems, improving translation quality while maintaining inference efficiency (Zeng et al., 2024). Overall, existing studies demonstrate that LLMs have strong multilingual translation capabilities, while data composition, prompt design, and parameter adaptation strategies remain important factors affecting their multilingual translation performance. Meanwhile, differences in representations across languages and cross-lingual parameter interference remain important challenges limiting the effectiveness of multilingual adaptation of LLMs.

2.2 Multilingual Parameter Adaptation and Direction-Specific Modeling

PEFT has become increasingly important for adapting LLMs to multilingual tasks at low computational cost (Han et al., 2024). Prefix-Tuning (Li and Liang, 2021b), for example, freezes the backbone model while learning a small number of continuous prefix parameters, which are injected into the attention mechanisms of Transformer layers as additional key/value representations, thereby enabling downstream task adaptation with a relatively small number of trainable parameters. Prompt-based approaches have also demonstrated effective adaptation capabilities, with several studies showing that carefully designed prompts can substantially improve the translation quality of LLMs (Zhang et al., 2023; Gao et al., 2024; Yamada, 2023; Khoboko et al., 2025).

Beyond prompt-level adaptation, module-based PEFT methods have also been extensively studied. Su et al. (2024) conducted a comprehensive comparison of 15 PEFT architectures supported by AdapterHub (Pfeiffer et al., 2020), showing that Houlsby+Inversion Adapter achieves strong overall performance in low-resource NMT. More recently, language-aware adaptation mechanisms have been proposed to alleviate cross-lingual parameter interference, including language-specific neuron routing strategies (Zhu et al., 2024b) and sparse mixture-of-experts architectures such as Lingual-SMoE, which selectively activate subsets of parameters for different languages (Zhao et al., 2024).

In contrast to these approaches, DiET performs multilingual adaptation through translation-direction-specific parameterization. Given a source–target language pair, DiET treats it as an explicit translation direction identifier and deterministically selects the corresponding direction-specific encoder from a set of encoders outside the frozen LLM backbone. The selected encoder generates input-dependent continuous prefix representations based on the current source sentence, which are then injected into the frozen LLM for translation adaptation. In this way, DiET combines direction-specific parameterization with input-dependent representation construction, providing relatively independent external adaptation spaces for different translation directions.

3 Methodology

As illustrated in Figure 2, we propose DiET, a direction-specific encoder tuning framework for multilingual machine translation with large language models. With the LLM backbone parameters frozen, DiET introduces a set of direction-specific encoders outside the backbone model and deterministically selects the corresponding encoder based on the known source–target language pair. The selected encoder re-encodes the source sentence to generate input-dependent continuous prefix representations, which are then fed into the frozen LLM together with the translation instruction and source text, thereby enabling adaptation parameter isolation across different translation directions.

DiET aims to provide independent adaptation spaces for different translation directions through direction-specific parameterization, thereby reducing parameter interference across directions. Unlike conventional fine-tuning methods, DiET does not modify the architecture or parameters of the backbone model. Instead, it introduces an external, non-invasive mechanism that dynamically generates direction-specific prefix representations to guide the translation behavior of the LLM.

Specifically, DiET introduces independent direction-specific encoders, maintaining a separate encoder parameter space for each translation direction. For each input sample, the corresponding direction-specific encoder dynamically generates continuous representations based on the source sentence. These prefix representations are injected into the frozen backbone in a non-invasive manner, providing the frozen LLM with input representations associated with the current translation direction while preserving the original parameter space of the backbone model. The overall architecture of DiET is illustrated in Figure 2.

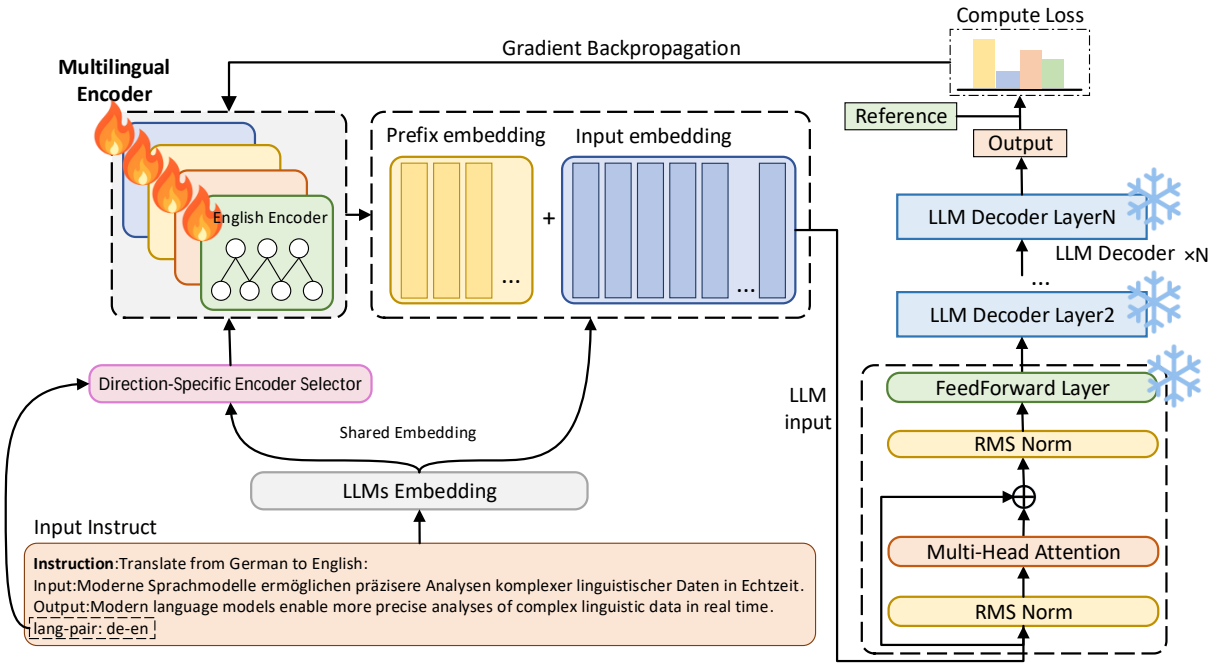


Figure 2: Overview of the proposed DiET framework.

3.1 Direction-specific Encoder

Due to substantial differences in syntax, morphology, and word order across languages, multilingual machine translation with LLMs faces inherent challenges arising from linguistic diversity. When methods such as LoRA or Prefix-Tuning jointly optimize multilingual data using shared parameters, these linguistic differences can lead to negative parameter interference, semantic confusion, and performance degradation in certain translation directions.

To address the limitations caused by cross-lingual parameter interference and achieve more robust multilingual adaptation, DiET introduces direction-specific encoders to explicitly decouple representations at the encoder level. For each translation direction, DiET maintains an independent Transformer-based direction-specific encoder. Given an input sample, DiET deterministically selects the corresponding encoder from a predefined encoder set based on its known translation-direction label.

Let s and t denote the source and target languages, respectively. A translation direction is defined as

$$d = (s, t). \tag{1}$$

For a predefined set of translation directions $\mathcal{D}$, DiET establishes a corresponding encoder for each direction $d \in \mathcal{D}$:

$$\mathcal{E} = \{E_d \mid d \in \mathcal{D}\}. \tag{2}$$

Given a translation sample (x, d), where x denotes the source-language text and $d = (s, t)$ denotes its corresponding translation direction, DiET deterministically selects the corresponding encoder from the encoder set $\mathcal{E}$ based on the known direction identifier d:

$$E_d = \text{Select}(\mathcal{E}, d). \tag{3}$$

The direction label d serves as a fixed index into the encoder set $\mathcal{E}$, selecting the corresponding encoder E_d. Samples with the same direction label use the same encoder, while samples from different translation directions update their respective encoder parameters independently.

Given a source sentence x, its input representation is first obtained through the shared input embedding layer of the frozen LLM:

$$H_{\text{src}} = \text{Embed}_{\text{LLM}}(x), \tag{4}$$

where

$$H_{\text{src}} \in R^{B \times S \times d_{\text{llm}}}, \tag{5}$$

B denotes the batch size, S denotes the source sequence length, and d_{llm} denotes the input embedding dimension of the LLM. Importantly, only

the source sentence representation H_{src} is used during the encoding process; tokens from the translation instruction are not provided to the direction-specific encoder. This design allows the encoder to primarily model the linguistic and semantic information of the source sentence itself.

3.2 Source Re-encoding and Continuous Prefix Generation

For the selected direction-specific encoder E_d, DiET first maps the source representation from the LLM embedding space to the encoder hidden space. If the LLM input embedding dimension d_{llm} differs from the encoder hidden dimension d_{enc}, a trainable projection matrix $\mathbf{W}_{\text{in}}$ is used for dimensionality mapping:

$$H_{\text{enc}}^{(0)} = H_{\text{src}}\mathbf{W}_{\text{in}}, \tag{6}$$

where

$$\mathbf{W}_{\text{in}} \in R^{d_{\text{llm}} \times d_{\text{enc}}}, \qquad H_{\text{enc}}^{(0)} \in R^{B \times S \times d_{\text{enc}}}. \tag{7}$$

The resulting $H_{\text{enc}}^{(0)}$ is then fed into the selected direction-specific Transformer encoder. For the l-th encoder layer, the self-attention module is formulated as:

$$Q^{(l)} = H_{\text{enc}}^{(l-1)} W_Q^{(l)}, \tag{8}$$

$$K^{(l)} = H_{\text{enc}}^{(l-1)} W_K^{(l)}, \tag{9}$$

$$V^{(l)} = H_{\text{enc}}^{(l-1)} W_V^{(l)}, \tag{10}$$

where $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$ denote the query, key, and value projection matrices of the l-th self-attention layer, respectively. The self-attention output is computed as:

$$A^{(l)} = \text{softmax}\left(\frac{Q^{(l)}(K^{(l)})^\top}{\sqrt{d_h}}\right) V^{(l)}, \tag{11}$$

where d_h denotes the dimensionality of each attention head. After self-attention, the feed-forward network, and the corresponding residual and normalization operations, the output of the l-th layer is denoted as:

$$H_{\text{enc}}^{(l)} = \text{TransformerLayer}_l\left(H_{\text{enc}}^{(l-1)}\right). \tag{12}$$

After L_d layers of the direction-specific Transformer encoder, the final re-encoded source representation is obtained as:

$$H_{\text{enc}} = H_{\text{enc}}^{(L_d)}. \tag{13}$$

The depth L_d can be configured for each translation direction, allowing different directions to employ encoders with different depths.

To map the encoder hidden representation back to the input embedding space of the frozen LLM, DiET employs a direction-specific output projection matrix $\mathbf{W}_{\text{out}}$:

$$P_d = H_{\text{enc}}\mathbf{W}_{\text{out}}, \tag{14}$$

where

$$\mathbf{W}_{\text{out}} \in R^{d_{\text{enc}} \times d_{\text{llm}}}, \qquad P_d \in R^{B \times S \times d_{\text{llm}}}. \tag{15}$$

P_d is the input-dependent continuous prefix representation generated by DiET for the current translation direction and source sentence. Unlike directly optimized static continuous prompts, P_d varies with the source sentence and can therefore further model the linguistic and semantic characteristics of the current input within the direction-specific parameter space.

3.3 Continuous Prefix Injection

After obtaining the direction-specific continuous prefix representation P_d, DiET concatenates it with the translation instruction and source-text embeddings and feeds the resulting sequence into the frozen LLM.

Let r denote the translation instruction and I its number of tokens. The shared input embedding layer of the frozen LLM is used to obtain:

$$H_{\text{ins}} = \text{Embed}_{\text{LLM}}(r), \tag{16}$$

and the source representation H_{src}. Their dimensions are:

$$H_{\text{ins}} \in R^{B \times I \times d_{\text{llm}}}, \qquad H_{\text{src}} \in R^{B \times S \times d_{\text{llm}}}. \tag{17}$$

The continuous prefix, translation instruction, and source text are then concatenated along the sequence dimension to form the final input representation of the LLM:

$$H_{\text{LLM}} = \text{Concat}\left[P_d,\, H_{\text{ins}},\, H_{\text{src}}\right]. \tag{18}$$

Since P_d has the same sequence length S as the source sentence, the resulting input representation has the shape:

$$H_{\text{LLM}} \in R^{B \times (2S+I) \times d_{\text{llm}}}. \tag{19}$$

The corresponding attention mask is also extended along the sequence dimension, allowing

the frozen LLM to perform standard autoregressive computation over the complete input sequence containing the continuous prefix, translation instruction, and source text.

During training, all LLM parameters remain frozen, and only the direction-specific encoders and their input and output projection parameters are optimized. For a translation direction d, its training samples update only the corresponding E_d and associated projection parameters, without directly updating the encoders of other translation directions. Thus, different translation directions maintain independent encoder parameter spaces while sharing a unified language modeling foundation through the frozen LLM's input embedding space and pretrained knowledge.

Given a training sample (x, y, d), where x denotes the source text, y denotes the target translation, and d denotes the translation direction, the model computes the conditional probability of the target sequence using the frozen LLM and adopts the standard autoregressive negative log-likelihood as the training objective:

$$\mathcal{L} = -\sum_{t=1}^{|y|} \log p_{\Theta_{\mathrm{LLM}},\Phi_d}\left(y_t \mid y_{<t}, H_{\mathrm{LLM}}\right), \quad (20)$$

where Θ_{LLM} denotes the frozen LLM parameters, and Φ_d denotes the trainable parameters associated with translation direction d, including the direction-specific encoder and its projection parameters. During training, only Φ_d is updated, while Θ_{LLM} remains fixed.

4 Experiments

We conduct experiments on Qwen3-4B and LLaMA-3.2-3B-Instruct across multiple translation directions, and compare DiET with a series of representative strong baseline methods.

4.1 Datasets

To systematically evaluate the effectiveness of **DiET** in MMT, we conduct experiments on several publicly available and widely used parallel corpora, including NTREX-128 [1] (Federmann et al., 2022), WMT18, and the ALT subset of the OPUS parallel corpus [2]. These datasets cover six languages: English, Burmese, Chinese, German, Russian, and Czech, spanning the Indo-European and Sino-Tibetan language families. The substantial differences in linguistic structure and resource availability provide a suitable basis for evaluating translation performance under diverse linguistic characteristics and resource conditions.

We construct bidirectional translation tasks between English and five other languages, including Burmese, Chinese, German, Russian, and Czech. Specifically, the tasks cover translation from these five languages into English and from English into the corresponding target languages. The selected language pairs exhibit substantial differences in data resource sizes and cover high-, medium-, and low-resource settings, providing an appropriate testbed for evaluating the effectiveness of DiET under multilingual data imbalance.

The training data are primarily collected from NTREX-128, WMT18, and ALT, comprising approximately 820K parallel sentence pairs in total. Considering the substantial imbalance in data resources across language pairs, we apply a unified data cleaning and filtering procedure and randomly sample high-quality parallel sentence pairs for training. As shown in Table 1, the sizes of the training sets vary across language pairs, ranging from 10K to 100K sentence pairs.

For evaluation, we use FLORES+ [3] (NLLB Team et al., 2024; Goyal et al., 2022b), which extends FLORES-200 to more than 200 languages and is one of the widely adopted benchmarks for multilingual machine translation.

	en-my	en-zh	en-ru	en-cs	en-de
Train	10K	100K	100K	100K	100K
Test	1K	1K	1K	1K	1K

Table 1: Dataset statistics for the bilingual parallel corpora in the En→X translation direction. Given the symmetry of the dataset, the X→En splits are equivalent and thus omitted.

4.2 Model Settings

We implement the proposed method on two frozen backbone models: Qwen3-4B [4] (Yang et al., 2025; Team, 2025) and LLaMA-3.2-3B-Instruct [5]. Both models exhibit strong multilingual modeling and instruction-following capabilities.

[1] https://github.com/MicrosoftTranslator/NTREX

[2] https://opus.nlpl.eu/ALT/en&my/v20191206/ALT

[3] https://huggingface.co/datasets/openlanguagedata/flores_plus

[4] https://huggingface.co/Qwen/Qwen3-4B

[5] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

To isolate the effect of DiET, we keep all backbone parameters frozen throughout training and update only the parameters of the direction-specific encoders and their associated modules. Each direction-specific encoder is implemented as a Transformer encoder, whose hidden dimension matches the embedding dimension of the corresponding backbone model, i.e., 2560 for Qwen3-4B and 3072 for LLaMA-3.2-3B-Instruct. For a given backbone, this hidden dimension is fixed across translation directions, while the encoder depth L_d may vary. The source sequence length S varies across input samples.

Since different language pairs may differ in semantic complexity, different translation directions may benefit from different encoder depths. For example, English→Burmese can benefit from increasing encoder depth, whereas Czech→English and Russian→English achieve their best observed performance with one-layer and three-layer encoders, respectively. Therefore, we maintain an independent encoder for each language pair and translation direction, allowing each encoder to process variable-length inputs to preserve fine-grained semantic information.

For multi-head self-attention, following previous empirical studies (Michel et al., 2019), we select the maximum number of attention heads that evenly divides the hidden dimension, with an upper bound of 16.

We train each model for 3 epochs with a learning rate of 2e-6, a batch size of 1, and a gradient clipping threshold of 5. During inference, the input to the frozen backbone consists of the task instruction, the continuous prefix representations generated by the direction-specific encoder, and the source sentence. This design provides translation-specific guidance for different translation directions without modifying the model's original instruction-following capability.

We use the fastText language identification model [6] (Joulin et al., 2016b,a) to validate the output language, particularly for languages that use similar writing systems. A translation result is considered valid when the language identification confidence exceeds 0.75.

Translation quality is evaluated using BLEU and COMET-22 [7]. BLEU measures the surface-level n-gram overlap between the generated translation and the reference translation, while COMET-22 provides a learned-model-based evaluation using reference translations and has been shown to exhibit higher agreement with human judgments.

In addition, we use XLM-RoBERTa-large [8] as a pretrained language model for the relevant language processing and evaluation procedures.

To evaluate the effectiveness of DiET, we further conduct a series of comparison experiments against various parameter fine-tuning methods and model adaptation strategies:

- **LoRA** (Hu et al., 2022b): Introduces lightweight low-rank updates into the decoder and shares the same LoRA configuration across all language pairs.
- **P-tuning** (Liu et al., 2022): Prompt-based tuning that optimizes continuous prompt embeddings for downstream tasks.
- **Adapter** (Houlsby et al., 2019): Inserts Adapter modules into the backbone model and updates only the Adapter parameters while keeping the remaining model parameters frozen.
- **LSLo** (Cao et al., 2024): Employs language-specific sparse expert (Sparse MoE) layers to selectively fine-tune parameters for different languages.
- **LoRA-Flow** (Wang et al., 2024): Introduces a flow-based optimization mechanism on top of LoRA to better model cross-lingual knowledge transfer and alleviate parameter interference.

4.3 Main Results

Table 2 reports the BLEU and COMET-22 scores for five translation directions (de→en, cs→en, zh→en, ru→en, and my→en).

For Qwen3-4B, the frozen backbone already exhibits strong multilingual translation capabilities. DiET achieves the highest average BLEU score of 31.82, outperforming the frozen backbone and the evaluated PEFT baselines. Its average COMET-22 score is 80.70, compared with 80.78 for the frozen backbone and 80.66 for LoRA.

For LLaMA-3.2-3B-Instruct, the overall translation performance is relatively lower, particularly for Chinese and Burmese translation directions,

[6] https://huggingface.co/julien-c/fasttext-language-id

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

[8] https://huggingface.co/FacebookAI/xlm-roberta-large

Methods	BLEU						COMET					
	de-en	cs-en	zh-en	ru-en	my-en	Avg.	de-en	cs-en	zh-en	ru-en	my-en	Avg.
Qwen3-4B												
Base model	38.17	34.34	30.44	33.68	17.59	30.84	83.23	81.87	82.27	80.67	75.87	80.78
P-tuning	38.26	35.10	30.75	33.10	17.07	30.86	83.25	82.00	81.21	80.48	76.00	80.59
Adapter	38.42	35.85	30.86	34.30	17.98	31.48	82.46	82.51	82.59	80.39	75.28	80.65
LoRA	38.50	35.80	31.16	34.32	18.32	31.62	82.55	81.33	82.56	81.11	75.76	80.66
LSLo	38.05	35.49	30.86	33.83	18.06	31.26	83.19	81.93	82.26	81.31	74.74	80.69
LoRA-Flow	38.04	35.50	30.90	33.98	17.21	31.13	82.62	82.75	82.34	80.65	74.81	80.63
Ours DiET	**38.72**	**35.92**	**31.52**	**34.46**	**18.47**	**31.82**	**82.98**	**80.87**	**82.39**	**81.32**	**75.93**	**80.70**
Llama-3.2-3B-Instruct												
Base model	33.29	28.65	23.76	29.08	7.91	24.54	81.44	77.94	79.56	78.62	52.96	74.10
P-tuning	33.78	27.99	23.77	28.91	7.05	24.30	80.25	77.21	79.42	78.25	51.92	73.41
Adapter	33.35	28.01	23.75	30.14	9.64	24.98	80.46	78.22	79.82	78.86	58.26	75.12
LoRA	33.50	28.39	23.82	29.64	10.01	25.07	81.82	79.26	79.97	78.39	61.90	76.27
LSLo	33.15	28.63	23.17	29.72	10.06	24.95	80.41	79.47	80.18	79.65	59.77	75.90
LoRA-Flow	33.53	28.05	23.74	29.93	10.11	25.07	81.13	78.97	79.81	79.46	58.87	75.65
Ours DiET	**37.92**	**33.35**	**25.78**	**32.80**	**12.94**	**28.56**	**82.80**	**80.59**	**80.99**	**79.95**	**61.53**	**77.17**

Table 2: COMET and BLEU scores on the FLORES+ devtest set for X→En translation. Results are reported for two base models (Qwen3-4B and Llama-3.2-3B-Instruct) under different adaptation methods. Bold numbers indicate the results of our DiET.

which may be related to differences in the model's multilingual modeling capability. Under this setting, DiET achieves an improvement of more than 2 BLEU on most translation directions, with average gains of 4.02 BLEU and 3.07 COMET-22, respectively. The En→X experimental results in Appendix Table 6 further support this trend. Overall, these results demonstrate that DiET can reduce cross-lingual negative interference through direction-specific external adaptation without modifying the backbone model parameters, providing stable and highly generalizable multilingual translation improvements across different backbone models.

4.4 Ablation Study

To analyze the contributions of different components in DiET, we conduct ablation experiments on representative high-, medium-, and low-resource language pairs. All experimental variants are trained and evaluated under the same experimental conditions, differing only in the enabled components. The results are presented in Table 3.

We consider the following configurations for the ablation study. **LoRA** applies lightweight low-rank adaptation to the decoder and shares the same configuration across all languages. **Prefix-Only** optimizes only the prefix representations while keeping the backbone model fully frozen. **Shared-Encoder** shares a single encoder across all language pairs and translation directions and uses this encoder to generate prefix representations. **DiET** adopts a direction-specific encoder architecture, assigning an independent encoder to each translation direction. Under both Shared-Encoder and DiET, the backbone model parameters remain frozen, and only the parameters of the corresponding encoders are updated.

These configurations enable a systematic comparison of different parameter-efficient adaptation strategies, including prefix-based optimization and cross-lingual parameter sharing, and allow us to analyze their effects on multilingual translation performance.

The ablation analysis mainly focuses on the comparison between Shared-Encoder and DiET to examine the effect of translation-direction-specific parameterization. Compared with Prefix-Only, the main difference of Shared-Encoder lies in how the prefix representations are constructed: it replaces directly optimized static parameters with representations dynamically generated by a shared encoder based on the input. However, since all languages are still constrained within the same shared representation space, the performance improvement of Shared-Encoder is relatively limited, particularly for medium- and low-resource language pairs with greater linguistic heterogeneity and data imbalance.

In contrast, DiET exhibits stronger and more stable performance across different resource conditions. By assigning an independent encoder to each translation direction, DiET reduces cross-lingual in-

terference during training, allowing each language pair to learn more specialized representations. This design is particularly beneficial for low-resource languages, as shared adaptation parameters are more susceptible to being dominated by data from high-resource languages.

Overall, the ablation results support the effectiveness of combining translation-direction-specific parameterization with input-dependent representation construction. The comparison between Shared-Encoder and DiET further demonstrates the importance of effectively isolating direction-specific representations when adapting LLMs to multilingual machine translation.

Methods	my-en	cs-en	de-en	zh-en	ru-en
Qwen3-4B					
Base model	17.59	34.34	38.17	30.44	33.68
+Lora	18.32	35.80	38.50	31.16	34.32
+Prefix-only	12.67	27.10	30.23	27.75	27.10
+Shared-Encoder	15.32	35.00	38.10	30.05	34.88
+DiET	18.47	35.92	38.72	31.52	34.46

Table 3: BLEU-based ablation results on X→En translation with Qwen3-4B under different adaptation settings.

5 Analysis

5.1 Effect of DiET Encoder Depth

We investigate the effect of the depth of the DiET direction-specific encoder on translation quality while keeping the encoder hidden dimension and the backbone model's input embedding dimension unchanged. Due to computational constraints, we restrict the search range of encoder depth to 1–6 Transformer layers. Figure 3 presents the results for X→En, while Figure 5 in the appendix presents the results for En→X. The results show that the optimal encoder depth varies across language pairs and resource conditions.

For Qwen3-4B, the results show that the effect of encoder depth varies across translation directions. In particular, low-resource translation directions generally benefit more from deeper encoder configurations. For Qwen3-4B, for example, English→Burmese achieves 16.33 BLEU and 15.54 BLEU with one-layer and two-layer encoders, respectively, while its performance increases substantially to 19.23 BLEU with a six-layer encoder. In contrast, high-resource translation directions can often achieve strong performance with shallower encoders, while overly deep encoders may lead to overfitting. For example,

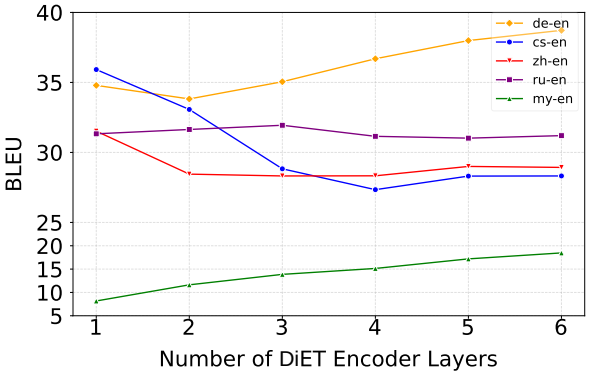


Figure 3: Effect of different DiET encoder depths on BLEU scores across multiple language pairs under the X→En translation setting with the Qwen3-4B backbone.

English→Russian already achieves 29.88 BLEU with a one-layer encoder. However, this pattern is not universal, as some high-resource language pairs can also benefit from increasing the encoder depth. These observations suggest that the optimal encoder depth in DiET is related not only to data resource availability but may also be influenced by the linguistic characteristics of different translation directions.

Overall, these results indicate that using a uniform encoder depth is not necessarily optimal. Instead, selecting an appropriate encoder depth according to the characteristics of each translation direction can lead to better translation performance.

5.2 Effect of backbone input length

We further investigate the effect of the maximum input sequence length of the backbone model on translation performance. In DiET, the model input includes not only the task instruction and source sentence, but also the continuous prefix representations generated by the direction-specific encoder. Therefore, the input capacity is an important factor affecting model performance.

Due to computational constraints, we evaluate maximum input sequence lengths of 800, 2560, 3840, and 5120 tokens. Figure 4 reports BLEU scores for X→En, while Figure 6 in the appendix reports BLEU changes relative to the 800-token setting for En→X. The results show that increasing the maximum input length does not consistently improve translation performance, and its effect is highly dependent on the language pair. An insufficient maximum input length may cause information loss due to truncation, thereby degrading translation quality. In contrast, an excessively large length limit may introduce a large number of padding

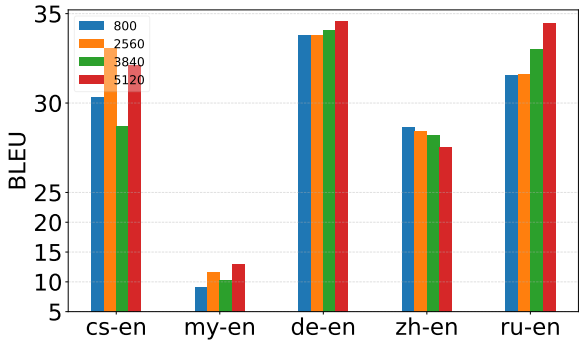


Figure 4: Effect of different backbone model maximum input lengths on BLEU scores across multiple language pairs under the X→En translation setting with the Qwen3-4B backbone.

tokens, causing attention to be distributed over a large number of sparse tokens, which may reduce the model's focus on useful information and negatively affect some language pairs.

Taking Qwen3-4B as an example, with a two-layer encoder, English→Burmese translation performance improves substantially as the maximum input length increases from 800 to 5120 tokens, whereas Czech→English achieves its best performance at a moderate length of 2560 tokens.

It should be noted that these conclusions are empirical observations based on a limited experimental setting. In practical applications, we recommend selecting an input length that sufficiently preserves the source sentence and its corresponding continuous prefix representations while avoiding excessive padding, and adjusting it according to the characteristics of different language pairs to balance translation performance and computational efficiency.

6 Conclusion

We propose **DiET**, a direction-specific encoder tuning framework for MMT. DiET adopts a frozen-backbone adaptation strategy, keeping all LLM backbone parameters frozen without modifying its original architecture. DiET introduces a set of direction-specific encoders to generate input-dependent continuous prefix representations from the source sentence, which provide guidance for effectively adapting the model to MMT tasks. By assigning an independent encoder to each translation direction, DiET achieves parameter decoupling across translation directions and reduces cross-directional parameter interference during multilingual joint training. Experiments under different resource conditions on two representative LLM backbones, Qwen3-4B and LLaMA-3.2-3B-Instruct, demonstrate that DiET improves BLEU across the evaluated translation directions, with more pronounced COMET gains on LLaMA-3.2-3B-Instruct.

Limitations

Although DiET provides an effective adaptation approach for multilingual machine translation, it still has several limitations. First, DiET maintains an independent encoder for each translation direction, which may introduce additional parameter and memory overhead when scaling to a large number of languages. Second, the quality of the prefix representations depends on the capacity and architectural design of the encoders, and may still be limited by insufficient training signals in extremely low-resource settings. Finally, DiET currently focuses primarily on sentence-level machine translation, and extending it to document-level or context-aware translation remains an open direction for future research.

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, 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.

Zhe Cao, Zhi Qu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Exploring intrinsic language-specific subspaces in fine-tuning multilingual neural machine translation. *arXiv preprint arXiv:2409.05224*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025a. 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.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025b. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

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.

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*.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. *arXiv preprint arXiv:2305.07310*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022a. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022b. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR) Workshop*. ArXiv:2106.09685.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022b. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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

Pitso Walter Khoboko, Vukosi Marivate, and Joseph Sefara. 2025. Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, 20:100649.

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. 2021a. 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, Online.

Xiang Lisa Li and Percy Liang. 2021b. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

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.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hanqing Wang, Bowen Ping, Shuo Wang, Xu Han, Yun Chen, Zhiyuan Liu, and Maosong Sun. 2024. Lora-flow: Dynamic lora fusion for large language models in generative tasks. *arXiv preprint arXiv:2402.11455*.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into chatgpt's customizability. *arXiv preprint arXiv:2308.01391*.

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*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Improving machine translation with large language models: A preliminary study with cooperative decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13275–13288, Bangkok, Thailand.

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*.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024a. LANDeRMT: Dectecting 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.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024b. Landermt: Detecting and routing language-aware neurons for selectively finetuning llms to machine translation. *arXiv preprint arXiv:2409.19523*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024c. 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.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024d. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

A Appendix

Table 4 reports the BLEU-based ablation results of Qwen3-4B on the En→X translation directions. Although the frozen backbone model performs competitively on high-resource translation directions, its performance degrades substantially on low-resource target languages such as Burmese. LoRA mainly benefits high-resource translation directions. In contrast, Prefix-only leads to a systematic performance degradation, indicating its limited capacity for target-language-specific modeling. Shared-Encoder achieves some improvements

over Prefix-only but remains constrained by a unified representation space. In comparison, DiET achieves the best performance across all translation directions, with particularly substantial improvements on low-resource languages while maintaining competitive results on high-resource language pairs. These results indicate that, with the backbone model parameters frozen, direction-specific parameterization can effectively improve multilingual adaptation.

Methods	**en-my**	**en-cs**	**en-de**	**en-zh**	**en-ru**
			Qwen3-4B		
Base model	15.65	24.00	31.88	27.36	27.78
+Lora	13.80	25.84	33.95	28.39	29.11
+Prefix-only	10.82	24.36	31.75	27.45	26.92
+Shared-Encoder	12.00	24.60	33.18	26.94	27.77
+DiET	19.23	26.23	34.64	29.86	29.88

Table 4: BLEU-based ablation results on En→X translation with Qwen3-4B under different adaptation settings. Base model indicates no fine-tuning, while other variants differ in parameter update scope and language-specific adaptation.

Table 5 presents the BLEU-based ablation results of Llama-3.2-3B-Instruct on both X→En and En→X translation directions. Since the frozen backbone model has relatively limited multilingual translation capabilities, its performance is relatively weak on most language pairs, particularly on low-resource translation directions such as English→Burmese.

LoRA and Shared-Encoder provide moderate performance gains over the baseline. Prefix-only performs poorly across all settings, indicating that globally shared prefix parameters have difficulty capturing translation characteristics specific to different languages.

In contrast, DiET achieves the best performance across all language pairs and translation directions, with particularly substantial improvements under low-resource settings. These results indicate that even when the backbone model has relatively limited multilingual capabilities, DiET can effectively reduce cross-lingual parameter interference and achieve robust direction-specific multilingual adaptation with the backbone model parameters frozen.

Figure 5 shows the optimal DiET encoder depth for different language pairs under the En→X translation setting with the Qwen3-4B backbone. Some translation directions achieve their best performance with relatively shallow encoders, whereas others benefit from increasing the encoder depth.

Methods	**my-en**	**cs-en**	**de-en**	**zh-en**	**ru-en**
			Llama-3.2-3B-Instruct		
Base model	7.91	28.65	33.29	23.76	29.08
+Lora	10.01	28.39	33.50	23.82	29.64
+Prefix-only	5.28	25.40	29.15	20.00	28.93
+Shared-Encoder	8.08	28.67	32.48	24.73	29.25
+DiET	12.94	33.35	37.92	25.78	32.80
Methods	**en-my**	**en-cs**	**en-de**	**en-zh**	**en-ru**
Base model	1.98	18.32	27.00	20.53	20.92
+Lora	2.29	19.14	27.65	21.61	20.97
+Prefix-only	2.01	19.27	25.53	20.21	20.46
+Shared-Encoder	2.78	19.05	27.13	19.89	21.20
+DiET	4.09	22.56	33.04	22.96	24.71

Table 5: BLEU-based ablation results for Llama-3.2-3B-Instruct under different adaptation strategies in both X→En and En→X translation directions. The base model is evaluated without fine-tuning, while the remaining variants differ in their parameter update scope and direction-specific adaptation mechanisms.

This variation cannot be explained solely by differences in data resource sizes and may instead be related to the intrinsic characteristics of the languages, such as language family, syntactic structure, and semantic complexity. These results suggest that there is no single optimal encoder depth applicable to all translation directions.

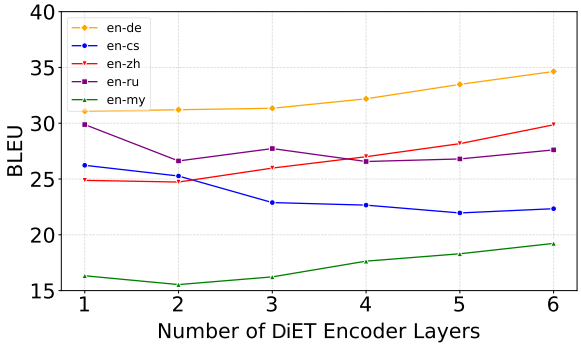


Figure 5: Effect of DiET encoder depth on BLEU scores across multiple language pairs for the Qwen3-4B backbone in the En→X translation setting, with encoder depth varying from 1 to 6 layers.

Figure 6 shows the BLEU improvement relative to the 800-token baseline when using a two-layer DiET encoder for En→X translation with Qwen3-4B.

The results show that the optimal maximum input length varies across languages. Specifically, English→Chinese achieves the highest BLEU at 800 tokens, but exhibits a continuous upward trend from 3840 to 5120 tokens, suggesting that further increasing the input length may still provide

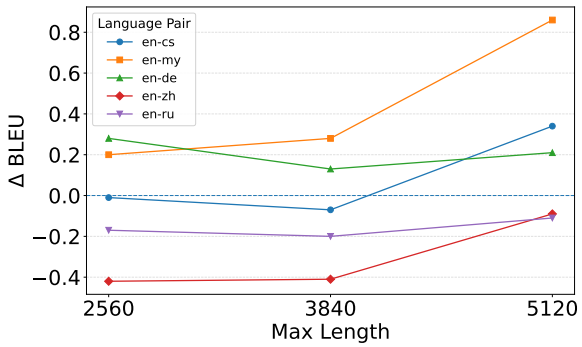


Figure 6: BLEU improvements over the 800-token baseline for En→X translation using the Qwen3-4B model. Performance gains are computed as the BLEU difference between maximum input lengths of 2560, 3840, and 5120 tokens and the 800-token setting. The x-axis denotes the maximum input length, and the y-axis represents the corresponding BLEU improvement.

potential benefits. In contrast, English→German achieves its best performance at 2560 tokens but exhibits some fluctuations at longer input lengths. These results indicate that the optimal maximum input length is language-dependent and is also affected by computational resource constraints.

We further investigate translation quality in the En→X translation directions to examine whether the effectiveness of DiET generalizes to translation scenarios where English is the source language and other languages are the target languages. We evaluate the translation results on the FLORES+ devtest set using BLEU and COMET, with the results summarized in Table 6.

This setting covers both high-resource and low-resource target languages, providing a complementary perspective on multilingual model adaptation. Overall, the results are consistent with the observations in the X→En setting: DiET achieves stable and competitive performance improvements across different target languages and backbone models, indicating that its effectiveness is not limited to a specific translation direction or target-language distribution.

This advantage is particularly pronounced on Llama-3.2-3B-Instruct, whose baseline translation performance is relatively weak. With the introduction of DiET, the model achieves average improvements of 3.72 BLEU and 6.13 COMET over the frozen backbone model, substantially outperforming other parameter-efficient adaptation methods. These results indicate that even under low-resource conditions and with relatively weak baseline performance, DiET can effectively reduce cross-lingual parameter interference and improve overall multilingual translation quality.

In summary, the En→X translation results further demonstrate that the advantages of DiET generalize across different translation directions, resource conditions, and target-language distributions, highlighting its robustness and practical potential as a frozen-backbone, direction-specific external adaptation method.

Methods	BLEU						COMET					
	en-de	**en-cs**	**en-zh**	**en-ru**	**en-my**	**Avg.**	**en-de**	**en-cs**	**en-zh**	**en-ru**	**en-my**	**Avg.**
Qwen3-4B												
Base model	31.88	24.00	27.36	27.78	15.65	25.33	80.58	79.27	83.08	82.94	60.39	77.25
P-tuning	31.81	25.38	28.34	27.16	12.37	25.01	81.09	80.71	82.29	84.95	61.23	78.05
Adapter	33.15	25.33	28.36	28.72	13.98	25.91	81.46	80.51	83.59	83.39	61.15	78.02
LoRA	33.95	25.84	28.39	29.11	13.80	26.22	82.00	81.86	83.25	84.39	62.01	78.70
LSLo	32.07	25.49	27.87	29.34	14.06	25.77	81.19	80.23	82.26	84.10	60.74	77.70
LoRA-Flow	33.03	25.50	28.20	28.98	14.21	25.98	81.62	80.75	82.34	83.65	61.81	78.03
Ours DiET	**34.64**	**26.23**	**29.86**	**29.88**	**19.23**	**27.97**	**81.66**	**80.35**	**83.19**	**84.52**	**61.74**	**78.29**
Llama-3.2-3B-Instruct												
Base model	27.00	18.32	20.53	20.92	1.98	17.75	76.07	59.53	79.94	68.14	43.00	65.34
P-tuning	26.78	18.49	20.77	19.91	2.00	17.59	75.79	64.21	79.42	70.25	42.92	66.52
Adapter	27.10	19.01	21.75	20.14	2.34	18.07	76.46	65.22	80.18	71.86	43.26	67.40
LoRA	27.65	19.14	21.61	20.97	2.29	18.33	76.78	65.34	80.48	71.09	43.10	67.36
LSLo	27.15	19.63	21.17	20.72	2.06	18.15	76.85	66.47	79.82	71.12	42.97	67.45
LoRA-Flow	27.53	19.05	20.74	20.93	2.11	18.07	76.13	64.97	79.81	71.06	43.87	67.17
Ours DiET	**33.04**	**22.56**	**22.96**	**24.71**	**4.09**	**21.47**	**80.81**	**72.73**	**81.00**	**77.04**	**45.77**	**71.47**

Table 6: BLEU and COMET evaluation results on the FLORES+ devtest set for En→X translation. Results are reported across high-resource and low-resource target languages using Qwen3-4B and Llama-3.2-3B-Instruct as backbone models. Bold numbers indicate the results of our DiET.