SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 150104:1–150104:17
https://doi.org/10.1007/s11432-025-4914-1

Special Topic: Large Multimodal Models

MNAFT: modality neuron-aware fine-tuning of multimodal large language models for image translation

Bo LI[2,5], Ningyuan DENG[3], Tianyu DONG[1], Shaobo WANG[4], Shaolin ZHU[1*] & Lijie WEN[2*]

[1]*School of Computer Science and Technology, Tianjin University, Tianjin 300072, China*
[2]*School of Software, Tsinghua University, Beijing 100084, China*
[3]*School of Information Resource Management, Renmin University of China, Beijing 100872, China*
[4]*School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China*
[5]*Baidu Inc., Beijing 100193, China*

Received 31 March 2025/Revised 29 September 2025/Accepted 23 March 2026/Published online 17 April 2026

Abstract Multimodal large language models (MLLMs) have shown impressive capabilities, yet they often struggle to effectively capture the fine-grained textual information within images crucial for accurate image translation. This often leads to a modality gap between visual text inputs and textual inputs/outputs for image translation. Existing methods, primarily relying on instruction fine-tuning, risk parameter redundancy of pre-trained knowledge, hindering generalization performance. To address this, we introduce modality neuron-aware fine-tuning (MNAFT), a novel approach that takes advantage of the specialized roles of individual neurons within MLLMs for enhanced image translation. MNAFT identifies language-agnostic and language-specific neurons in both vision and language modules through an instruction-driven activation analysis, evaluating their importance in various translation tasks. We then perform selective fine-tuning, updating only the parameters of language-specific and language-agnostic neurons within the selected layers relevant to the target task, while preserving the knowledge encoded in other neurons and layers. Our extensive experiments on multiple benchmarks demonstrate that MNAFT significantly outperforms state-of-the-art image translation methods, including cascaded models, standard full fine-tuning, and parameter-efficient tuning techniques. Furthermore, we provide comprehensive analysis, including visualizations of neuron activations and clustering patterns, to offer insights into the roles of different neuron groups in mediating cross-modal understanding and facilitating accurate language-specific translation.

Keywords vision-language models, multilingual image translation, large language models

Citation Li B, Deng N Y, Dong T Y, et al. MNAFT: modality neuron-aware fine-tuning of multimodal large language models for image translation. Sci China Inf Sci, 2026, 69(5): 150104, https://doi.org/10.1007/s11432-025-4914-1

1 Introduction

Multimodal large language models (MLLMs) have recently revolutionized the landscape of artificial intelligence, demonstrating remarkable capabilities in tasks that require a unified understanding of both visual and textual information, such as visual question answering, image captioning, and visual reasoning [1–7]. These models, typically built by integrating a pre-trained visual encoder with a powerful large language model (LLM) via a connector module, represent a significant advance towards building general-purpose AI systems. However, a crucial frontier for MLLMs lies in their application to complex, real-world tasks that demand fine-grained understanding and manipulation of multimodal information. Image translation (IT), which aims to translate text embedded in images from the source language to the target language, is precisely such a challenge [8–12].

The ability of IT has profound implications for the global accessibility of visual content, cross-cultural communication, and international commerce, making it a critical area of research [13]. Traditional approaches to IT, often employed in systems like Google Translate's Instant Camera and Google Lens, rely on a cascaded pipeline of optical character recognition (OCR) followed by machine translation (MT). This approach suffers from error propagation (OCR inaccuracies that impact MT), computational inefficiency (due to sequential processing) and a lack of holistic contextual understanding, since the OCR and MT components operate independently [13–15].

End-to-end (E2E) image translation models offer a more direct and potentially more accurate solution [8,16–18], but adapting MLLMs to this task effectively requires addressing a fundamental modality gap. Visual encoders within MLLMs are typically pre-trained on vast image-text datasets using contrastive learning [2,6,19–23]. While

* Corresponding author (email: zhushaolin@tju.edu.cn, wenlj@tsinghua.edu.cn)

this pre-training is effective for general visual-text understanding, it may not be optimal for capturing the subtle, nuanced characteristics of multilingual text embedded within images, which are critical for high-fidelity image translation. This discrepancy between the visual-text input representation and the desired textual output leads to suboptimal translation accuracy and fluency. Models such as InternVL2 [20], LLaVA-NeXT [21], LLaMA3.2 [24], and Qwen2.5-VL [25], despite their impressive general capabilities, are demonstrably affected by this limitation.

The predominant paradigm for adapting MLLMs to specific tasks is instruction fine-tuning, where the pre-trained model is further trained on task-specific data, often with instructions guiding the model's output. Fine-tuning MLLMs for IT [15] is intuitive, but the prevailing practice, exemplified by models such as LLaVA-NeXT [21], involves uniformly updating all model parameters. This approach neglects the functional specialization of individual neurons within MLLMs. Neurons are not homogeneous; they develop distinct roles in processing different modalities and languages. Uniform updates risk disrupting pre-trained knowledge and are suboptimal for specialized tasks like image translation. Although previous work explores network analysis [26], it lacks a practical and targeted method of fine-tuning based on neuron specialization to improve multimodal task performance. A more refined approach, explicitly leveraging neuron-specific roles, is essential to unlock the full potential of MLLMs for image translation.

To address this issue, we introduce modality neuron-aware fine-tuning (MNAFT), a novel and principled method specifically designed to optimize MLLM for image translation by leveraging the functional specialization of individual neurons. MNAFT is based on the key insight that different neurons within the vast network of an MLLM develop different roles: some become specialized in processing specific languages, others in handling visual features, and still others in bridging the gap between modalities. By accurately identifying and selectively targeting these neurons during fine-tuning, we can achieve significantly improved translation performance while preserving the general capabilities of the MLLM. MNAFT consists of two core stages. (i) We introduce a robust methodology with instruction-driven, using both activation patterns and gradient information, to identify language-specific and modality-shared neurons within the visual and textual processing pathways of the MLLM. This creates a precise "functional map" of neuron specialization within the network. (ii) Guided by the identification of neurons, we implement a selective fine-tuning strategy. During fine-tuning on target language image translation data, only the parameters of language-specific and language-agnostic neurons within the selected layers relevant to the target task are updated. The remaining neurons are frozen, preventing disruption of pre-trained knowledge and avoiding parameter redundancy. This targeted approach maximizes fine-tuning efficiency and minimizes unintended consequences.

We perform a rigorous evaluation of MNAFT across a comprehensive suite of image translation benchmarks, encompassing diverse datasets and translation tasks. Our experimental results demonstrate that MNAFT substantially outperforms state-of-the-art baselines, establishing a new performance benchmark for the field. Moreover, we provide an in-depth analysis, including novel visualizations, that elucidates the functional roles of different neuron groups in mediating cross-modal understanding and achieving accurate, language-specific translation. This analysis provides unprecedented insights into the inner workings of MLLMs, contributing significantly to our fundamental understanding of these powerful models.

Our contributions can be summarized as follows.

- We introduce MNAFT, the first modality neuron-aware fine-tuning method specifically designed to optimize MLLM for image translation, achieving superior performance and efficiency.
- We present a detailed analysis with novel visualizations that reveal the functional specialization of neurons within MLLMs during cross-modal and language-specific processing, significantly advancing our understanding of these complex models.
- We demonstrate, through comprehensive experiments, that MNAFT achieves state-of-the-art results in image translation, setting a new benchmark for the field.

2 Related work

2.1 Image translation

IT aims to translate texts embedded in images from the source language to the target language [9]. Its wide range of applications makes it a valuable field of research [11,12]. Current image translation systems can be divided into two paradigms: cascaded and end-to-end approaches. Cascaded approaches employ an OCR to extract text from the input image, followed by a separate neural machine translation (NMT) for translation [27–29]. However, this approach suffers from error propagation, massive parameters, and complexity of deployment [13,14]. Eventually, end-to-end IT that integrates OCR and MT modules into a single model has attracted much attention [16]. Ref. [17]

applied multi-task learning to this task, where NMT and OCR are jointly trained. Furthermore, Ref. [18] applied knowledge distillation to effectively distill the knowledge of OCR and NMT into end-to-end IT. Ref. [8] explored an end-to-end IT with an aligner and a regularizer to reduce the modality gap, and Ref. [9] introduced an IT model with multimodal codebooks to reduce the impact of OCR errors. Ref. [13] used a target text decoder and an image tokenizer to alleviate the language alignment burden and improve performance by transforming long pixel sequences into shorter visual token sequences. Recently, MLLMs [2, 5–7, 21, 23, 24, 30] have demonstrated impressive performance in various tasks such as visual question answering, visual understanding, and reasoning. These solutions normally follow to utilize the visual encoder to encode visual features and utilize the connector module to project visual tokens into the word embedding space of the LLM. However, the visual encoder (e.g., CLIP), which is primarily pre-trained on image-text pairs with contrastive learning. Therefore, it is an intuitive solution to fine-tune MLLM to enhance performance on the IT task [15].

2.2 Multimodal large language model instruction tuning

Instruction tuning has significantly improved the generalization capability of MLLMs on various tasks [1,15,31,32]. However, standard full fine-tuning updates a large number of weights in all intermediate layers and the pre-trained LLM, leading to parameter redundancy and high computational costs [33, 34]. Therefore, parameter-efficient fine-tuning (PEFT) methods have been proposed [35–38]. Among them, low-rank adaptation (LoRA) [35] and its variants such as DORA [39] have become widely accepted as PEFT methods, where the fine-tuning of models is performed by updating a small number of injected adaptation parameters. However, in multimodal instruction tuning, traditional PEFT methods often suffer from parameter redundancy, as they rely on fitting a limited number of common parameters to perform different tasks, severely affecting the transferability between previously learned datasets [40–43]. Moreover, most existing PEFT methods focus on single modalities and neglect the crucial role of multimodal features in fine-tuning [44–46]. For example, some methods typically freeze the visual encoder and only fine-tune the connector layers and the LLM component, limiting the model's ability to fully utilize multimodal information. To address these limitations, MixLoRA [47] builds on LoRA by dynamically constructing low-rank adaptation matrices tailored to the unique requirements of each input, with the goal of mitigating task interference. M^2PT [48] facilitates cross-modal feature extraction and matching of cross-modal features by injecting visual and textual prompts into the visual and textual processors, respectively. In contrast to the aforementioned methods, our proposed MNAFT method explicitly considers the effects of both cross-modal interactions and language-specific neuron behavior on image translation. MNAFT first identifies language-specific and general neurons within the model by analyzing their activation patterns and gradient information during cross-modal interactions. During fine-tuning for a target language, MNAFT then selectively updates only the relevant cross-modal and language-specific neurons while freezing others, thus preserving knowledge about other languages and modalities. This targeted strategy minimizes interference with the translation capabilities of other languages and improves generalization performance.

3 Preliminaries

3.1 Task definition

Formally, we define the task of IT using a dataset $\mathcal{D} = \{(v_i, s_i, t_i)\}_{i=1}^{N}$, where v_i denotes the input image with the text of the source language. s_i denotes the text string of the source language extracted from v_i, which serves as auxiliary information for neuron analysis. t_i denotes the corresponding target language translation of s_i. The goal of IT is to generate the optimal target translation t_i directly from the input image v_i. During training, we can express the loss function as follows:

$$\mathcal{L}_{\mathrm{IT}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(t_i|v_i;\theta), \tag{1}$$

where θ denotes the parameters of the MLLM.

3.2 Taylor expansion

The Taylor expansion provides a polynomial approximation to a function [8, 49, 50]. We use it to analyze how the impact of removing a neuron affects the loss of the model. For a given neuron i with activation h_i, we examine the

effect of setting $h_i = 0$ on the loss function $\mathcal{L}(H, h_i)$, where H denotes the activations of all other neurons. The Taylor expansion of $\mathcal{L}(H, h_i)$ by one point a is as follows:

$$\mathcal{L}(H, h_i) = \sum_{n=0}^{\infty} \frac{\mathcal{L}^{(n)}(H, a)}{n!}(h_i - a)^n, \tag{2}$$

where $\mathcal{L}^{(n)}(H, a)$ is the n-th derivative of $\mathcal{L}$ with respect to h_i, evaluated with respect to a. For practical purposes, we use the first-order approximation:

$$\mathcal{L}(H, h_i) \approx \mathcal{L}(H, a) + \frac{\partial \mathcal{L}(H, a)}{\partial h_i}(h_i - a). \tag{3}$$

To analyze the impact of removing the neuron i, we set $a = 0$:

$$\mathcal{L}(H, h_i) \approx \mathcal{L}(H, 0) + \frac{\partial \mathcal{L}(H, 0)}{\partial h_i} h_i. \tag{4}$$

The change in loss $\Delta\mathcal{L}(h_i)$ due to the distance of the neuron is then approximated as follows:

$$\Delta\mathcal{L}(h_i) = \mathcal{L}(H, 0) - \mathcal{L}(H, h_i) \approx -\frac{\partial \mathcal{L}(H, 0)}{\partial h_i} h_i. \tag{5}$$

The magnitude of this change, $|\Delta\mathcal{L}(h_i)|$, reflects the importance of the neuron. We define the importance value $\Theta_{\mathrm{TE}}(i)$ as

$$\Theta_{\mathrm{TE}}(i) = \left| \frac{\partial \mathcal{L}(H, 0)}{\partial h_i} h_i \right|, \tag{6}$$

where $|h_i|$ denotes the magnitude of activation of the neuron, reflecting its contribution to the output of the network. $\left|\frac{\partial \mathcal{L}(H,0)}{\partial h_i}\right|$ denotes the sensitivity of the loss to changes in neuronal activation, evaluated at $h_i = 0$. A higher $\Theta_{\mathrm{TE}}(i)$ denotes a greater importance. This first-order Taylor expansion provides an efficient and effective way to estimate the importance of neurons for our selective fine-tuning strategy.

4 Method

Figure 1 illustrates the key components of our MNAFT methodology. We begin by evaluating the importance of neurons within both the vision layers and language layers across different languages using a novel instruction-driven approach. This yields neuron importance scores that reflect their contribution to the image translation task. We then select the most influential neurons within key layers of both modules and categorize them into language-agnostic (general) and language-specific groups. Finally, we perform selective fine-tuning by updating only the specific neurons relevant to the target task and the general neurons in the same layer to obtain the encoded general knowledge. This targeted approach mitigates parameter redundancy and parameter interference.

4.1 Insight

Different neurons within MLLMs exhibit varying degrees of specialization for different languages and visual modalities, posing challenges for standard fine-tuning. Indiscriminate fine-tuning can lead to: (i) erosion of general knowledge essential for robust multimodal understanding, and (ii) interference between language-specific features. MNAFT addresses this by employing an instruction-driven approach to identify and prioritize relevant neurons for each language and task. Instead of random perturbations, we use targeted instructions to test the activation of neurons, such as the instruction "What is the text in the picture?", which activates neurons responsible for recognizing and processing text within an image, and the instruction "Translate the text from [Source Language] into [Target Language]. [Source Language]: [Source Text] [Target Language]:" activates neurons specialized in translating between specific language pairs.

We propose a method of selective neuron adaptation that uses Taylor expansion (TE)-based importance scores to identify and prioritize the most relevant neurons for each language. Using Taylor expansion, we can effectively evaluate the importance of each neuron in different image translation tasks. In this context, we define the awareness score, denoted as $\Phi(i)$, for each neuron with respect to a particular task. With this innovative approach, we can

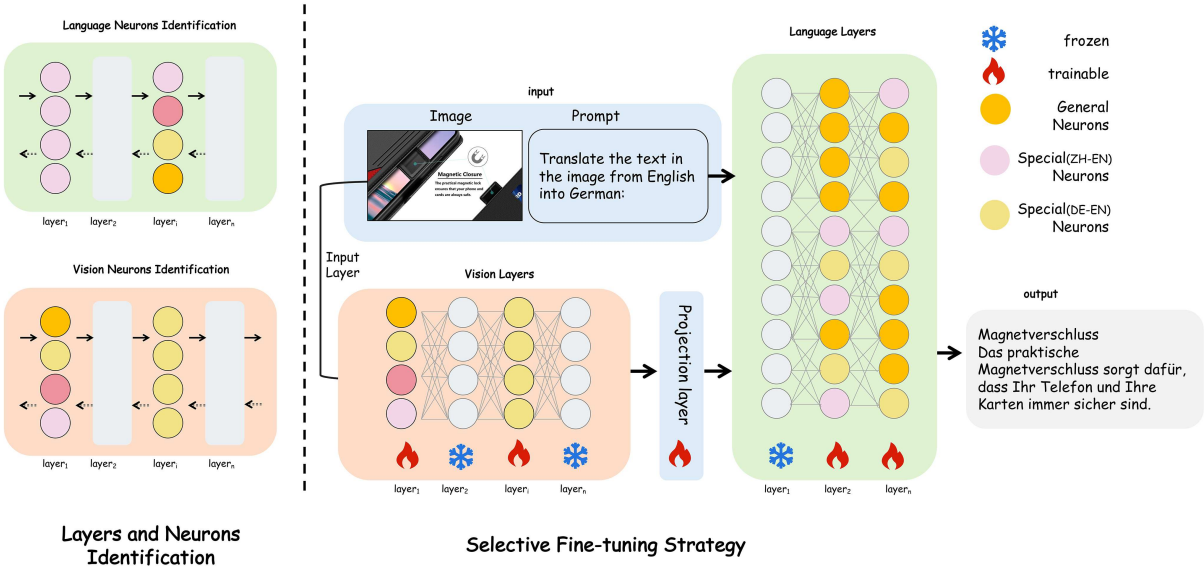

Figure 1 (Color online) The overview of MNAFT framework. We begin by evaluating the importance of neurons within both the vision layers and language layers across different languages using a novel instruction-driven approach. This yields neuron importance scores that reflect their contribution to the image translation task. We then select the most influential neurons within key layers of both modules and categorize them into language-agnostic (general) and language-specific groups. Finally, we perform selective fine-tuning by updating only the specific neurons relevant to the target task and the general neurons in the same layer to obtain the encoded general knowledge.

not only evaluate the importance of individual neurons but also distinguish between the neurons that are crucial for the processing of the entire task and the neurons that exert a particularly strong influence within a specific task.

$$\Phi(i) = |\Delta L(h_i)|, \quad i \in L_j, \tag{7}$$

where L_j denotes the layer previously selected. h_i denotes the output generated by the neuron i, while $|\Delta L(h_i)|$ refers to the corresponding loss value induced by perturbing specific neurons. By computing $\Phi(i)$ across multiple tasks and language pairs, we construct a matrix of the importance of neurons. This matrix reveals which neurons are consistently important across tasks (language-agnostic) and which specialize in particular languages or tasks (language-specific).

To establish a clear and explicit correlation between the neuronal activation value and its effect on the loss function, we utilize the Taylor expansion, which can be expressed as follows:

$$|\Delta L(h_i)| = \left| \frac{\partial L}{\partial h_i} h_i \right|. \tag{8}$$

This refined approach allows us to systematically evaluate the importance of neurons and prioritize those neurons that significantly contribute to translation accuracy. Unlike conventional magnitude-based pruning methods, which often overlook the intricate relationships between neuron activations and linguistic relevance, our gradient-aware significance metric demonstrates superior performance. By focusing on the nuanced contributions of individual neurons, we can enhance the overall effectiveness of image translation systems, ensuring that they remain robust and accurate in diverse linguistic contexts.

4.2 Neurons identification

Following the scoring process, we perform a thorough analysis of the layers by aggregating the neuron scores. This aggregation allows us to identify neurons that exhibit significant relevance to the task at hand. Understanding these significant neurons is crucial for revealing the modular characteristics inherent in the neural network architecture, as it provides insights into how different layers contribute to overall performance. To quantify the relevance of each neuron within a specific layer, we define the following equation:

$$\Theta_m^{\mathrm{TE}}(i_l) = \frac{1}{T_m} \sum_t \left| \frac{\delta L(H, h_{li})}{\delta h_{li}} h_l^i \right|. \tag{9}$$

In this equation, m denotes the total number of neurons within each layer and $\Theta_m^{\mathrm{TE}}(i_l)$ serves as the relevance score for that specific layer. This score reflects the impact of each neuron on the overall loss function, providing a clear metric for assessing their importance. To enhance the model's ability to focus on specific languages, particularly those that are less represented or possess unique linguistic features, we implemented a normalization process across the layers. This normalization is crucial for several reasons. It helps mitigate the impact of inherent biases that may arise from the training data, allowing the model to achieve a more balanced understanding of different languages.

Next, we rank the layers according to their calculated relevance scores, organizing them from highest to lowest importance. This ranking process is essential for selecting the most significant modules within both the vision layers and the language layers, allowing us to focus our optimization efforts on where they will be most effective. The selection of the top layers can be expressed mathematically as follows:

$$L_{\text{vision}} = \arg\max_{\text{top}_l} \{D_1, D_2, D_{k_{\text{vision}}}\}, \tag{10}$$

$$L_{\text{LLM}} = \arg\max_{\text{top}_l} \{D_1, D_2, D_{k_{\text{LLM}}}\}, \tag{11}$$

where L_{vision} and L_{LLM} represent the selected layers of the vision encoder and the LLM, respectively, while $D_1, D_2, D_{k_{\text{vision}}}$ and $D_1, D_2, D_{k_{\text{LLM}}}$ denote the relevance scores for the respective layers. We selected L_{vision} and L_{LLM} as the most critical layers, with l_{vision} and l_{LLM} serving as hyperparameters. This structured approach enables us to effectively identify and select the most impactful modules for further optimization.

After identifying the important layers, we proceed to rank the neurons within each layer based on their scores derived from the Taylor expansion. This ranking process is crucial, as it allows us to pinpoint the most influential neurons that contribute significantly to the overall task performance. To analyze a single layer effectively, we sort the neurons according to their variance, which provides insight into their individual contributions:

$$\lambda(i) = \text{sort}(\sigma(X_i))\,\lfloor \epsilon \times p \rfloor, \quad i \in L, \tag{12}$$

where p denotes the total number of neurons in the i layer, while ϵ is a predefined ratio that helps establish a threshold for classification. Neurons exhibiting a variance in the linguistic awareness score below the estimated threshold $\lambda(i)$ are categorized as neurons of general language, indicating their greater applicability across multiple languages. In contrast, those that exceed this threshold are classified as specific language neurons, highlighting their specialized role in understanding particular linguistic features.

To further refine our approach, we aggregated the neurons into distinct collections customized to each pair of languages. This targeted aggregation allows us to optimize the model's performance by focusing on the unique characteristics and nuances of the languages involved. By leveraging the insights gained from this ranking and classification process, we can enhance the model's ability to process and generate language more effectively, ultimately leading to improved outcomes in various linguistic tasks.

4.3 Selective fine-tuning strategy

MNAFT employs a highly selective fine-tuning strategy targeting specific neurons within the identified layers of the vision and language modules. This approach preserves the general knowledge captured by the pre-trained model while enabling efficient adaptation to the target task. The selection process, guided by neuron importance scores and variance analysis, results in two key sets of neurons for each relevant layer l: language-agnostic neurons and language-specific neurons.

During fine-tuning, we freeze all parameters except the weights and biases of the selected language-specific neurons and the language-agnostic neurons within the identified layers. This crucial detail ensures that the general knowledge encoded in language-agnostic neurons is preserved and contributes to overall performance. Let W_l and b_l represent the weight matrix and the bias vector of layer l, respectively. For each layer l selected for fine-tuning, we apply a mask M_l to the gradients during backpropagation. This mask is defined as follows:

$$M_l[i] = \begin{cases} 1, & \text{if } i \in A_l \cup S_{l,t}, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where i indexes the neurons in the layer l. This mask effectively zeros out the gradients for all neurons except those in A_l and $S_{l,t}$. The update rule for the parameters of layer l becomes:

$$W_l^{t+1} = W_l^t - \alpha \cdot (\nabla_{W_l} L_t \odot M_l), \tag{14}$$

$$b_l^{t+1} = b_l^t - \alpha \cdot (\nabla_{b_l} L_t \odot M_l), \tag{15}$$

where W_l^{t+1} and b_l^{t+1} are the updated weight matrix and bias vector for layer l at time step $t+1$, W_l^t and b_l^t are the weight matrix and bias vector at time step t. α is the learning rate, L_t is the task-specific loss at time step t, $\nabla_{W_l} L_t$ and $\nabla_{b_l} L_t$ are the gradients of the loss with respect to the weights and biases of layer l, respectively, $\odot$ denotes element-wise multiplication. This selective application of gradients ensures that only the chosen language-specific and language-agnostic neurons are updated during fine-tuning, preserving general knowledge while adapting to the specific target task.

5 Experiments

We conducted extensive experiments on six tasks in four publicly available image translation datasets to demonstrate the effectiveness of our proposed MNAFT compared to existing end-to-end and cascaded baseline approaches. We also performed ablation studies to analyze the contribution of each component within MNAFT.

5.1 Setup

5.1.1 *Datasets*

We conducted comprehensive experiments on six tasks in four publicly available image translation datasets. There are two synthetic and two real datasets.

• **ECOIT** [8] is a large-scale image translation dataset in the e-commerce domain, containing product images automatically crawled from a Chinese e-commerce website[1)] paired with post-edited target translations (480k sentences with 3.64M source tokens).

• **IIMT** [13] utilizes the IWSLT14 [51] German to English dataset and the IWSLT17 [52] French to English dataset to synthesize paired images. In addition, the background color of the image is selected randomly and the resolution of the images is 512 × 512. It comprises 452230 instances.

• **MIT-10M** [15] is a large-scale parallel corpus of multilingual image translation with over 10M image-text pairs derived from real-world data, which has undergone extensive data cleaning and multilingual translation validation. It contains 840k images and 14-languages image-text pairs.

• **OPUS-MIT-5M**[2)] is constructed by randomly sampling 5M sentence pairs from the OPUS corpus. This image translation dataset contains 5 million sentence pairs and 20 language pairs.

For each task, we constructed a training set by 100k items from the source dataset. The evaluation was performed on a test set of 100 items per task. Notably, the neuron identification stage of MNAFT was also conducted using the test set. The neuron identification stage does not involve any parameter updates; it only computes importance scores based on activation magnitudes and gradient information to determine the structural selection of neurons and layers. This process is analogous to calibration-based pruning, where a small reference set is used to assess neuron saliency without optimizing model weights. Since no learning occurs during this stage, the use of the test set does not introduce evaluation bias or test-set leakage.

5.1.2 *Baselines*

In this work, we compared MNAFT with three categories of baseline methods and with six SOTA IT models.

Cascaded models. We used EasyOCR[3)] and PP-OCRv3 [53] for text extraction from images in the datasets, followed by NLLB-200 [54] for text translation. It is a machine translation model primarily intended for machine translation research, especially for low-resource languages. It allows single-sentence translation between 200 languages.

MLLM baseline models. We used the Qwen2.5-VL-3B [25] model and evaluated its performance using four strategies: text-only (text-only translation based on the text extracted from the images), zero-shot, one-shot, and chain-of-thought (CoT) [55] prompting. The prompt templates are shown in Figure 2.

Fine-tuning methods. We compared MNAFT with established fine-tuning techniques: full fine-tuning and four SOTA PEFT methods.

• Full fine-tuning. The default approach, where both the visual and the language layers are updated during fine-tuning.

1) https://www.taobao.com/.
2) https://huggingface.co/datasets/liboaccn/OPUS-MIT-5M.
3) https://github.com/JaidedAI/EasyOCR.

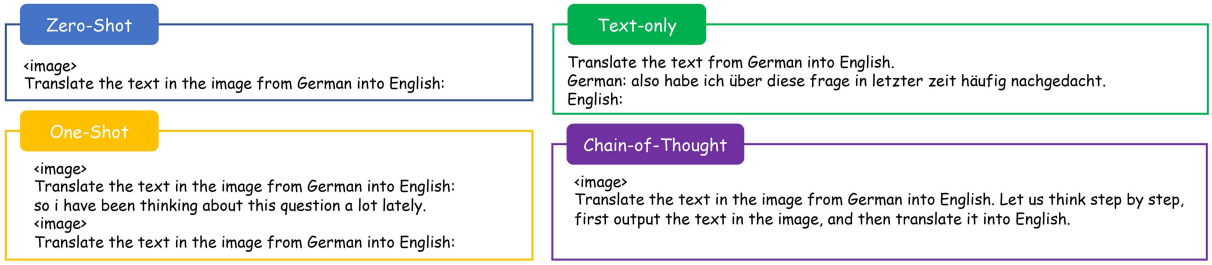

Figure 2 (Color online) Prompt templates. We used the Qwen2.5-VL-3B model and evaluated its performance using four strategies: text-only (text-only translation based on the text extracted from the images), zero-shot, one-shot, and CoT prompting.

• LoRA [35]. Efficiently fine-tunes large pre-trained models by using low-rank decomposition. It adds trainable low-rank matrices to the model's weight matrices, reducing the number of parameters that require fine-tuning. In our experiments, we configured LoRA with rank = 8 and applied it to all linear layers of the model (target_modules = all).

• DoRA [39]. Builds upon LoRA by decomposing pre-trained weights into "magnitude" and "direction" components for finer-grained control and learning. DoRA optimizes these magnitude and direction components separately.

• MixLoRA [47]. A novel approach that integrates multimodal instruction tuning with a conditional mixture of LoRA. Dynamically constructs low-rank adaptation matrices tailored to each input, instance by selecting their decomposition factors from two collections.

• M^2PT [48]. A multimodal prompt-tuning method for efficient instruction tuning of MLLMs. Introduces two sets of soft prompts: visual prompts and textual prompts, which are prefixed to the visual and instruction inputs, respectively.

SOTA IT models. As for the end-to-end model, we compare it with mainstream IT models.

• ItNet [16] is an end-to-end image translation system. It first pre-trains a standard transformer on a text-only parallel dataset. ResNet is used as an image encoder to encode latent semantic representations of images.

• PEIT [8] is an end-to-end image translation framework that bridges the modality gap with pre-trained models.

• Translatotron-V [13] is an end-to-end IT model consisting of four modules. In addition to an image encoder and an image decoder, it contains a target text decoder and an image tokenizer.

• UMTIT [56] first encodes the image using a vision transformer and then decodes the translation with a text transformer in an autoregressive manner.

• E2ETIT [17] builds a novel modal adapter that effectively fuses the OCR encoder and the MT decoder.

• DIMTDA [57] is document image machine translation with dynamic multi-pre-trained model assembly.

5.1.3 *Evaluation metrics*

We evaluated the performance of IT models on several dimensions, including semantic similarity, fluency, and accuracy. The evaluation metrics are as follows.

• BLEU [58]. It calculates the n-gram overlap between the candidate and reference translations.

• METEOR [59]. Taking into account synonyms and word order provides a more nuanced assessment of semantic similarity than BLEU.

5.1.4 *Implementation details*

The operating system that we use is CentOS release 7.5, and the programming language is Python 3.9.12. Our experiments were conducted on NVIDIA TESLA A100-80G GPU, the CUDA version is 12.2, and the deep learning framework is torch with version 2.1.0, torchvision with version 0.16.0 and Transformers with 4.45.0. We used the LlamaFactory [60] framework for all fine-tuning experiments.

5.2 Main result

Table 1 shows the image translation performance, measured by the METEOR and BLEU scores, for the six tasks and four datasets. Our method consistently shows superior performance compared to the baseline methods, highlighting its robustness and generalizability to different image translation scenarios.

Comparison to cascaded models. MNAFT significantly outperforms the traditional cascaded OCR + MT pipelines (EasyOCR_NLLB and PP-OCRv3_NLLB) on all tasks. In the task ECOIT (ZH-EN), for example, MNAFT achieves an METEOR score of 75.1 and is thus significantly higher than the scores of 13.7 and 13.1 of

Table 1 Quantitative comparison with three categories of baseline methods. We conducted a comprehensive evaluation of 11 models or methods (including cascaded, MLLM baseline models and fine-tuning method) using the test set from 6 language pairs on the 4 datasets. We highlight the best numbers in bold.

Model	ECOIT (ZH-EN)		IIMT (DE-EN)		IIMT (FR-EN)		MIT-10M (DE-EN)		MIT-10M (EN-DE)		OPUS-MIT-5M (EN-ZH)	
	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU
Cascaded models												
EasyOCR_NLLB	13.7	8.0	42.5	22.3	50.5	31.8	5.7	1.6	21.8	11.2	43.9	33.1
PP-OCRv3_NLLB	13.1	8.2	43.1	24.4	54.2	34.3	11.3	5.5	27.4	20.7	42.1	30.8
Baseline models (Qwen2.5-VL-3B)												
Text-only	51.9	35.6	48.4	29.8	56.7	39.9	63.2	50.7	63.6	55.1	60.9	45.2
Zero-shot	44.8	28.1	46.5	27.0	49.4	31.6	28.4	19.1	31.3	24.3	63.3	46.8
One-shot	49.0	31.9	39.1	19.6	43.3	24.6	33.3	16.1	30.7	21.8	52.2	40.1
Chain-of-thought	49.1	30.2	43.3	25.8	51.0	34.5	21.3	12.3	16.9	11.9	61.2	46.3
Fine-tuning methods												
Full fine-tuning	68.5	51.4	54.6	36.5	59.9	42.3	62.6	49.6	50.0	37.4	65.5	50.3
LoRA	62.2	45.7	56.9	38.4	60.5	44.5	57.1	40.6	46.1	35.6	66.7	51.8
DoRA	61.5	45.6	56.0	37.6	58.3	41.1	55.5	41.2	45.5	36.7	66.8	51.9
MixLoRA	61.3	44.2	61.6	34.2	61.8	41.3	55.1	40.9	45.4	36.1	66.6	53.4
M^2PT	63.7	46.2	61.0	34.0	64.3	43.2	58.0	46.8	49.7	36.3	66.1	53.1
MNAFT (ours)	**75.1**	**54.6**	**67.9**	**38.0**	**67.0**	**45.0**	**79.8**	**56.8**	**54.7**	**42.0**	**75.2**	**60.7**

the cascaded basic programs. This illustrates the limitations of relying on independent OCR and MT components, which are prone to error propagation and lack a holistic understanding of the visual and textual context.

Comparison to MLLM baselines. MNAFT also outperforms the Qwen2.5-VL-3B baseline on all zero-shot, one-shot and chain-of-thought prompting strategies. Importantly, MNAFT exceeds text-only on most tasks. This shows that it is able to integrate visual and textual information effectively, even when not relying on a real text. The improvement over zero-shot, one-shot, and chain-of-thought demonstrates the effectiveness of our method in utilizing training data for targeted performance improvements.

Comparison to fine-tuning methods. MNAFT outperforms all compared fine-tuning methods in different image translation datasets, including full fine-tuning, LoRA, DoRA, MixLoRA and M^2PT. This underlines the effectiveness of our neuron-aware fine-tuning strategy. For the MIT-10M (DE-EN) dataset, for example, MNAFT increases the METEOR score from 62.6 (full fine-tuning) to 79.8. By selectively updating only the most relevant neurons for the image translation task, MNAFT maximizes the benefits of fine-tuning while mitigating the risk of parameter redundancy.

5.3 Comparison with SOTA IT models

To further validate the effectiveness of MNAFT, we compare its performance with several state-of-the-art IT models. Table 2 shows the results for the ECOIT (ZH-EN) and IIMT (DE-EN) datasets. Our method achieves the best performance for both datasets and outperforms all six compared models in terms of METEOR and BLEU scores. In particular, for ECOIT (ZH-EN), MNAFT achieves an METEOR score of 75.1 and a BLEU score of 54.6, significantly outperforming the second-best model, Translatotron-V, which scores 73.1 and 52.6, respectively. In IIMT (DE-EN), MNAFT achieved an METEOR score of 67.9 and a BLEU score of 38.0 in this German-English translation task, a significant improvement over UMTIT's previous best results of 54.6 (METEOR) and 36.1 (BLEU).

These results show that our method is able to outperform existing specialized models for different language pairs. The superior performance can be attributed to the neuron-aware fine-tuning that effectively utilizes the multimodal capabilities of large language models while preserving the pre-trained knowledge and adapting to the specific nuances of the image translation task.

5.4 Ablation study

To investigate the contribution of the different components of MNAFT, we conducted an ablation study using the OPUS-MIT-5M (EN-ZH) and ECOIT (ZH-EN) datasets. The results are shown in Table 3. We analyzed the effects of fine-tuning different neuron groups and layers.

Effect of neuron type.

Table 2 Quantitative comparison with the SOTA IT models. We conducted a comprehensive evaluation of 6 models using the test set ECOIT (ZH-EN) and IIMT (DE-EN) dataset. We highlight the best numbers in bold.

	ECOIT (ZH-EN)		IIMT (DE-EN)	
	METEOR	BLEU	METEOR	BLEU
MNAFT (ours)	**75.1**	**54.6**	**67.9**	**38.0**
ItNet	61.1	39.3	48.9	27.3
PEIT	69.2	47.2	48.1	32.8
Translatotron-V	73.1	52.6	53.2	36.3
UMTIT	70.8	52.0	54.6	36.1
E2ETIT	46.1	31.5	32.1	21.9
DIMTDA	72.4	46.6	47.9	32.4

Table 3 Results of the ablation study comparing our method with different model variants on OPUS-MIT-5M (EN-ZH) and ECOIT (ZH-EN). We highlight the best numbers in bold.

	OPUS-MIT-5M (EN-ZH)		ECOIT (ZH-EN)	
	METEOR	BLEU	METEOR	BLEU
MNAFT (ours)	**75.2**	**60.7**	**75.1**	**54.6**
w/ General neurons FT	65.9	51.7	51.8	29.1
w/ Specific neurons FT	66.5	52.8	61.7	42.7
w/ ALL layers FT	67.2	53.3	61.6	45.8
w/ Language layers FT	66.5	51.0	58.7	42.3
w/ Vision layers FT	65.9	50.8	57.3	38.2

• General neurons FT. Fine-tuning only general neurons in both language and vision layers resulted in a significant drop in performance compared to MNAFT. This suggests that general neurons alone are not sufficient to achieve optimal performance in image translation. They likely capture broader features but lack the specific knowledge required for accurate translation.

• Specific neurons FT. Fine-tuning only specific neurons produced significantly better results than fine-tuning general neurons and approached the performance of MNAFT, especially on the ECOIT dataset. This highlights the crucial role of specific neurons in capturing the nuanced information necessary for image translation. These neurons are likely specialized for linguistic features of the target language or the visual representation of text.

Effect of layer selection.

• All layers FT (without neuron selection). Fine-tuning all neurons in both the language and visual layers, which is essentially MNAFT without the neuron selection mechanism, resulted in lower performance compared to the full MNAFT method. This demonstrates the importance of the neuron selection strategy to improve the effectiveness of the fine-tuning and prevent potential negative effects of updating irrelevant parameters.

• Language layers FT. Fine-tuning only the language layers, which includes both general and specific neurons, performed comparably well to the FT method for specific neurons in the OPUS-MIT-5M dataset, but worse in ECOIT. This suggests that the contribution of visual layer adaptation is more pronounced in some translation tasks.

• Vision layers FT. Fine-tuning only the visual layers resulted in the lowest performance among the ablation settings. This suggests that while adjusting the visual features is beneficial, the primary performance gain likely comes from adjusting the language layers that are directly responsible for producing the translated text.

This ablation study confirms the effectiveness of the core components of our method. The selection of specific neurons plays a crucial role in achieving optimal performance. While fine-tuning all layers provides some improvement over baseline MLLMs, the targeted approach of MNAFT shows superior performance. The results also suggest that the relative importance of matching vision and language layers may vary depending on the specific language pair and the characteristics of the dataset. The combination of specific neuron selection and joint vision-language fine-tuning leads to the best overall performance.

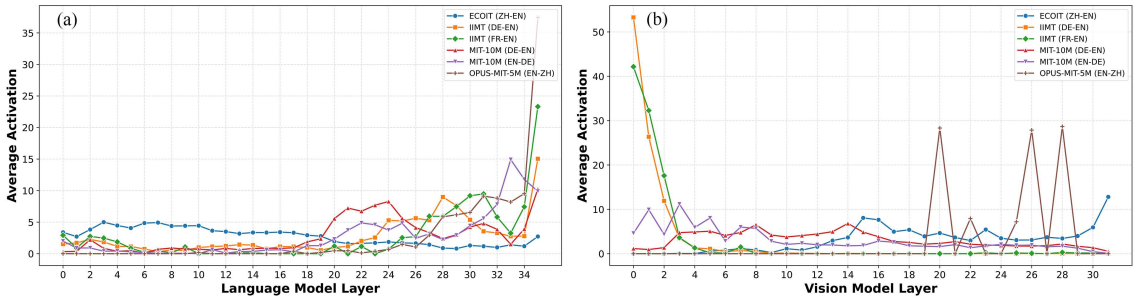

Figure 3 (Color online) Average activation. (a) Language model layer average activation; (b) vision model layer average activation.

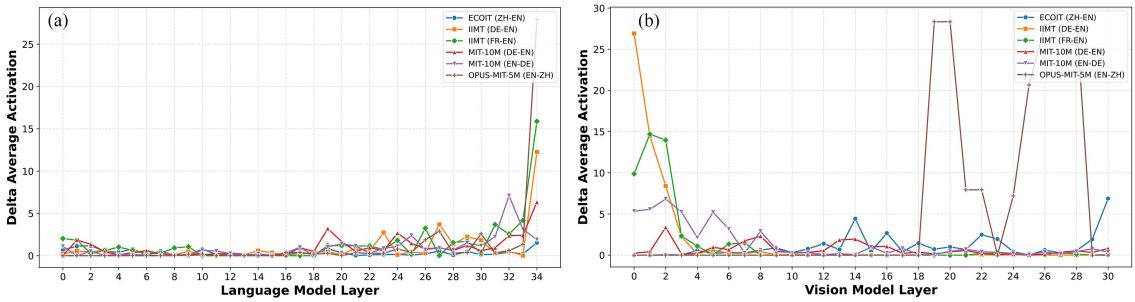

Figure 4 (Color online) Delta average activation. (a) Language model layer delta average activation; (b) vision model layer delta average activation.

6 Analysis

6.1 Neuron identification analysis

Figures 3 and 4 illustrate the average neuron activation and the average delta activation for each task for both the language model and vision model layers. Figure 3 shows which layers and neuron groups show higher overall activation in different tasks. For example, the OPUS-MIT-5M (EN-ZH) task shows significantly higher activation in specific layers of the vision model, suggesting that these layers are particularly sensitive to the visual features present in this dataset. Similarly, certain layers of the language model show higher activation for the ECOIT (ZH-EN) task. This difference in activation patterns between tasks is evidence of specialization within the model. Figure 4 shows the change in activation between successive layers. High delta values indicate layers where activation changes drastically, possibly indicating important processing steps or transitions between different levels of representation. The peaks observed for different tasks in different layers indicate task-specific processing within the model. For example, the sharp peak for the OPUS-MIT-5M dataset in the vision model suggests a significant shift in representation within that specific layer, possibly related to the processing of visual features unique to that task.

These visualizations in combination with our neuron selection method provide insight into how different parts of the model contribute to different image translation tasks. The identification of language-specific neurons and the observation of task-dependent activation patterns support the hypothesis that different neurons and layers are specialized to handle different aspects of the image translation process. This understanding motivates our neuron-based fine-aware fine-tuning strategy, which enables a more effective and targeted adaptation of the model to specific tasks.

6.2 Clustering analysis of general and specific neurons

To further investigate the role of general and specific neurons in capturing language and visual knowledge for image translation, we used t-SNE to visualize the representations learned by these neurons across different tasks. Figure 5 shows the t-SNE plots of neuron activations from the last layer of both vision and language modules.

Analysis of language model neurons (Figures 5(a) and (b)).

• **Specific neurons (a).** The t-SNE diagram of language-specific neurons shows clear clusters corresponding to different tasks of image translation. This clear separation indicates that these neurons are specialized in capturing task-specific linguistic features. The clusters for tasks with similar languages (e.g., DE-EN and FR-EN) appear to

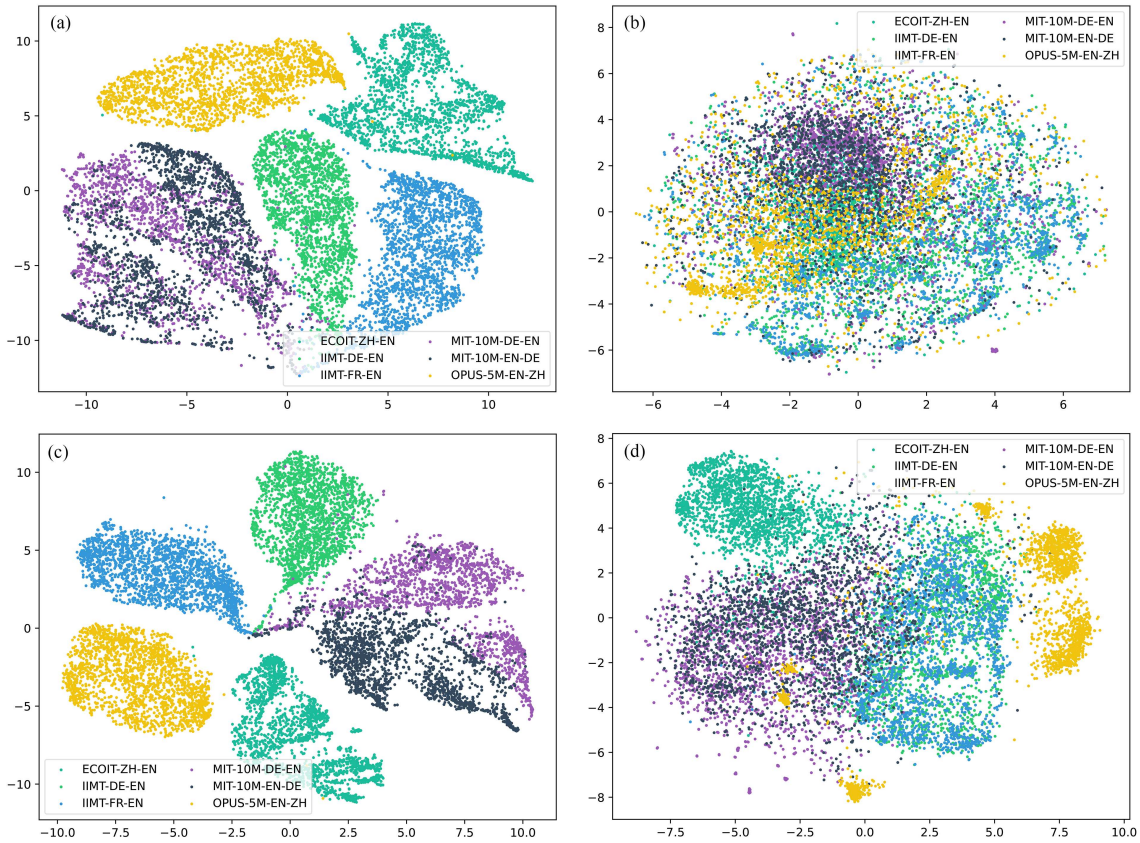

Figure 5 (Color online) Clustering of representations. (a) Language model layer specific neurons; (b) language model layer general neurons; (c) vision model layer specific neurons; (d) vision model layer general neurons.

be closer to each other than those with different languages (e.g., ZH-EN and EN-ZH), supporting our hypothesis that language-specific neurons learn linguistic nuances relevant to specific languages and translation directions.

• **General neurons (b).** In contrast, general neurons show a more mixed distribution. Although some loose groupings are evident, the lack of clear separation suggests that they capture more general task-related linguistic features, consistent with our expectation that they encode broader linguistic knowledge that applies to a wider range of languages.

Analysis of vision model neurons (Figures 5(c) and (d)).

• **Specific neurons (c).** Similar to the language model, the specific neurons of the vision model also show distinct, albeit less defined, task-related clusters, indicating specialization in task-relevant visual features. The less distinct clustering suggests that visual features are less strongly associated with individual languages than with linguistic features.

• **General neurons (d).** The general neurons in the vision model show a more diffuse distribution, suggesting that they capture general, task-related visual features, such as recognition of text regions or image layout. The less defined clustering supports the idea that general visual features in image translation are less task-specific than language-related features.

The t-SNE visualizations provide strong evidence for the specialization of specific neurons in the detection of task-related linguistic and visual features. This is consistent with the core motivation of MNAFT, which selectively tunes these neurons for improved task performance. The diffuse distribution of general neurons suggests that they play a role in capturing broader, task-related knowledge that is retained by our selective fine-tuning. These results validate the design choices of MNAFT and provide insights into the distinct roles of different neuron types in large multimodal models of image translation.

6.3 Results on Qwen2.5-VL-7B and LLaVA-NeXT

To evaluate the generalizability and scalability of our method, we conducted experiments with larger models: Qwen2.5-VL-7B and LLaVA-NeXT-LLaMA3 (8B). Figure 6 shows the consistent effectiveness of our method with

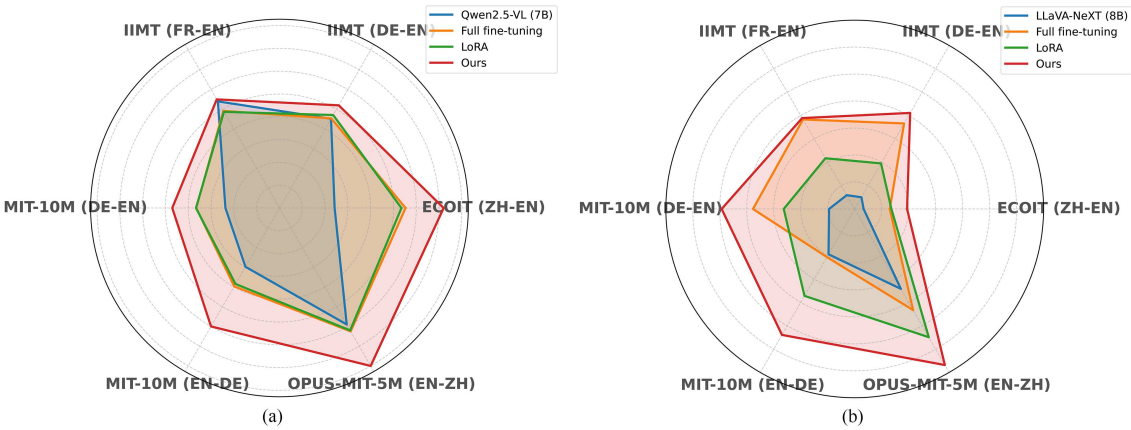

Figure 6 (Color online) Comparison of METEOR with six datasets for fine-tuning the Qwen2.5-VL-7B (a) and LLaVA-NeXT-LLaMA3 (8B) (b) using full fine-tuning, LoRA and MNAFT (ours).

different model architectures and sizes. The figure shows the METEOR scores obtained on the six image translation tasks for each model and fine-tuning method.

For the Qwen2.5-VL-7B model, MNAFT shows significant performance improvements compared to full fine-tuning and LoRA on all tasks. The figure shows that the MNAFT range significantly outperforms that of the other methods, indicating its overall superior performance. Although both full fine-tuning and LoRA improve the baseline performance of Qwen2.5-VL-7B (represented by the innermost hexagon), they fall well short of our method. This underscores the effectiveness of our neuronal fine-aware fine-tuning strategy in utilizing the increased capacity of the larger model. The experiments with LLaVA-NeXT further confirm the generalizability of our method. Although the overall METEOR score is lower for this model compared to Qwen2.5-VL-7B, MNAFT outperforms both full fine-tuning and LoRA on all tasks. The figure for LLaVA-NeXT shows a similar trend to the Qwen2.5-VL-7B experiments, with MNAFT covering a wider range than the other methods.

The consistent improvements observed in both Qwen2.5-VL-7B and LLaVA-NeXT demonstrate the robustness and scalability of MNAFT. Our neural fine-tuning strategy appears to be particularly effective in exploiting the increased capacity of larger models, leading to more significant gains compared to standard fine-tuning and LoRA. The figure clearly illustrates this superior performance on various translation tasks. These results suggest that our method offers a promising approach to maximizing the performance of MLLMs in IT tasks, regardless of the specific model architecture or scale.

6.4 Case study

To further illustrate the practical benefits of MNAFT, we present a qualitative analysis of its performance using specific examples. As shown in Figure 7, these case studies show how MNAFT leads to more accurate and contextually appropriate translations.

Case 1. This case shows an image with product packaging that contains both English text and a stylized logo. The base model (Qwen2.5-VL-3B) hallucinates additional text, "sicherheitshalber Ihre Smartphones und Karten immer", translating to "always your smartphones and cards for safety", which may be influenced by common product descriptions for cell phone cases. In contrast, MNAFT correctly translates the entire sentence as "Das praktische Magnetverschluss sorgt dafür, dass Ihr Telefon und Ihre Karten immer sicher sind". This demonstrates the robustness of MNAFT and its ability to capture the full meaning in the visual context.

Case 2. This case shows a promotional image with the text "100% Genuine Leather. High quality wallets". The base model provides a literal translation of "Hohe Qualität Geldbörsen" ("High quality wallets" for the second sentence). While this is grammatically correct, it misses the nuance of "High quality wallets" which is often understood as a product category. MNAFT correctly identifies and translates "Genuine Leather" as "Echtes Leder" and provides a more natural translation of "Hochwertige Geldbörsen" ("High-quality wallets"), demonstrating a deeper understanding of the implied meaning within the advertising context.

Case 3. This case is a headline image with the sentence "The Syrians created these routes and now they are being used against them". The base model translates this into Chinese, which means that "The Syrians created these routes and now they are using these routes". This translation loses the crucial context of the sentence, in which "they" in the second part of the sentence refer to the routes being used against the Syrians. MNAFT accurately

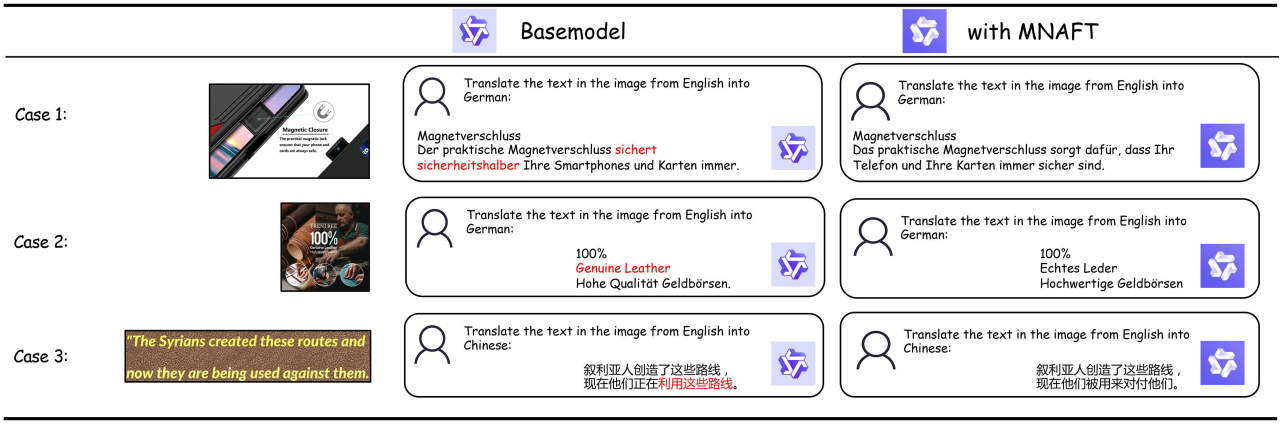

Figure 7 (Color online) Case study.

Table 4 Computational cost comparison on the ECOIT and OPUS-MIT-5M datasets. Time is reported in hours (h) and GPU memory in gigabytes (GB). Experiments are conducted on NVIDIA A100 (80 GB) GPUs. The neuron identification stage of MNAFT introduces only a negligible one-time cost (about 2 min for ECOIT and 3 min for OPUS-MIT-5M), which is not included in the table. All methods are trained for the same number of epochs and batch size.

Method	ECOIT		OPUS-MIT-5M	
	Time (h)	Memory (GB)	Time (h)	Memory (GB)
Full fine-tuning	9.2	94.3	12.6	127.7
LoRA	8.6	12.2	10.2	12.3
MNAFT (ours)	7.0	20.8	8.2	20.8

captures this nuance and translates it, which correctly reflects that the routes are used "against the Syrians". This demonstrates MNAFT's superior contextual understanding and disambiguation capabilities.

These case studies qualitatively demonstrate the advantages of MNAFT in tackling challenging scenarios in image translation. By selectively fine-tuning relevant neurons, MNAFT achieves a better balance between translation accuracy, visual fidelity, and contextual understanding, making it a promising approach for real-world image translation applications.

6.5 Computational cost analysis

To evaluate the training efficiency of our proposed MNAFT, we conducted a detailed comparison with full fine-tuning and LoRA in terms of wall-clock training time and peak GPU memory usage. The results on the ECOIT and OPUS-MIT-5M datasets are summarized in Table 4. The training pipeline of MNAFT comprises two stages: neuron identification and selective fine-tuning. The identification stage requires computing neuron importance scores once on a small scoring set. On ECOIT and OPUS-MIT-5M, this step is completed within 2 and 3 min, respectively, which is negligible compared to the overall training time (on the order of hours). Therefore, the efficiency of MNAFT is dominated by the subsequent fine-tuning phase.

Compared with full fine-tuning, MNAFT demonstrates substantial efficiency gains. In ECOIT, its fine-tuning is 24% faster (7.0 h vs. 9.2 h), while consuming only 22% of the GPU memory (20.8 GB vs. 94.3 GB). Similar improvements are observed in OPUS-MIT-5M. These results empirically validate our design of selectively updating a small subset of neurons to achieve significant reductions in training cost. The comparison with LoRA reveals an interesting trade-off. LoRA is the most memory-efficient method (about 12 GB), as it only introduces a small number of low-rank adaptation matrices. MNAFT uses moderately more memory (about 21 GB), because it must maintain optimizer states for a subset of the original parameters. Nevertheless, MNAFT achieves noticeably faster training: its fine-tuning phase is 18% faster on ECOIT and 20% faster on OPUS-MIT-5M. We attribute this advantage to LoRA's additional forward-pass operations (extra matrix multiplications), which introduce latency, while MNAFT retains the original architecture. Our method applies gradient masking during backpropagation, so optimizer updates are restricted to the selected neurons. Although gradients are still propagated for all activations in the current implementation, skipping parameter updates for frozen neurons reduces effective computation and leads to shorter training time.

Overall, MNAFT offers a favorable balance between time and memory: it is substantially more efficient than full fine-tuning and faster than LoRA while maintaining manageable memory overhead. However, we note that the

current implementation does not prune the gradient computation itself-masking is applied at the optimizer update stage. As a result, further efficiency gains may be possible with specialized implementations that support sparse backpropagation. We leave such optimizations for future work.

7 Broader applicability and future work

While the efficacy and efficiency of MNAFT have been thoroughly established within the domain of image translation, the underlying architectural insights and adaptive mechanisms are designed with broader applicability in mind. This section delves into the inherent versatility of MNAFT, exploring its potential to enhance performance across a wider spectrum of multimodal tasks and outlining future investigative pathways.

7.1 The foundation for generalized adaptability

MNAFT's operational premise rests on the observation that within expansive MLLMs, distinct neuronal populations assume specialized functions during their extensive pre-training. These specializations manifest in processing specific sensory inputs, linguistic structures, or intricate inter-modal relationships. This functional segregation is a universal characteristic of how complex neural architectures learn from diverse data, rather than being exclusive to image-to-text conversion. Our methodology, which integrates instruction-guided activation profiling with Taylor expansion-derived salience scores, provides a potent framework for the following.

- Identifying specific aggregations of neurons that demonstrate heightened activity or sensitivity when confronted with particular data types (e.g., visual attributes, grammatical constructs, emotional undertones) or in the context of specific assignments.
- Evaluating the precise impact these specialized neuronal units have on a given downstream objective, thereby enabling highly targeted parameter adjustments.
- By strategically preserving the weights of non-essential neurons and layers, MNAFT effectively counteracts common challenges such as catastrophic forgetting and the redundant updating of parameters-issues frequently encountered when adapting large, foundation models.

7.2 Envisioning applications beyond textual image conversion

We foresee substantial advantages in applying MNAFT's neuron-aware fine-tuning paradigm to a variety of other multimodal challenges.

- Crafting articulate descriptions for images demands both acute visual perception and fluent linguistic expression. MNAFT could focus on neurons responsible for detecting salient objects, dynamic actions, and their contextual relationships, alongside those crucial for constructing grammatically sound and contextually appropriate narratives. This would be particularly valuable for specialized domains (e.g., medical diagnostics, fashion commentary) or for generating multilingual captions.
- For visual question answering (VQA), it necessitates a comprehensive understanding of visual scenes coupled with the interpretation of natural language queries to formulate accurate text-based responses. MNAFT could be instrumental in isolating neurons dedicated to object recognition, property extraction, spatial reasoning, and query comprehension. For instance, specific neuron clusters might be optimized for discerning "color" attributes, while others excel at handling "quantity" inquiries. Targeted fine-tuning could sharpen the model's ability to respond to particular categories of visual questions with greater precision, bypassing the need for wholesale model retraining.
- In human-AI interaction, MLLMs must process visual information concurrently with natural language conversation. MNAFT could help customize the model for different dialog scenarios, such as locating objects referenced in discourse, inferring user intent from visual cues, or formulating visually grounded replies. Neuronal subsets specialized in dialog coherence or sentiment recognition could be precisely updated.

8 Conclusion

In this paper, we present MNAFT, a novel neuron-aware fine-tuning method to improve the performance of MLLMs for image translation. MNAFT exploits the insight that different neurons within MLLMs play different roles in the processing of multimodal information, some specialized for specific languages, and others capture general or cross-modal knowledge. By selectively fine-tuning only the most relevant neurons for a given image translation task, MNAFT maximizes the benefits of adaptation while mitigating the risk of parameter redundancy. Our extensive experiments with different datasets and language pairs have shown that MNAFT consistently outperforms existing

SOTA IT methods, including traditional cascade pipelines and various LLM fine-tuning strategies. The ablation studies confirmed the effectiveness of our neuron selection mechanism and the importance of joint vision-language adaptation. Further analyses using visualizations provided convincing evidence for the specialization of different neuron types within the model, supporting the underlying principles of MNAFT. The successful application of MNAFT to larger models such as Qwen2.5-VL-7B and LLaVA-NeXT demonstrated its scalability and generalizability in different MLLMs architectures. In the next step, we will investigate the transferability of the learned neuron specializations to different tasks and modalities. Applying MNAFT to other complex multimodal tasks, such as visual question answering or image captioning, could unlock significant performance improvements in these domains.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant Nos. 2024YFB3309702, 2023YFE0116400) and National Natural Science Foundation of China Youth Fund (Grant No. 62306210).

References

1 Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 49250–49267
2 Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023. 19730–19742
3 Wu J, Gan W, Chen Z, et al. Multimodal large language models: A survey. In: Proceedings of IEEE International Conference on Big Data, 2023. 2247–2256
4 Zhang D, Yu Y, Dong J, et al. Mm-llms: Recent advances in multimodal large language models. ArXiv:2401.13601
5 Gemini T. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv:2403.05530
6 Hong W, Wang W, Ding M, et al. Cogvlm2: Visual language models for image and video understanding. ArXiv:2408.16500
7 Xue L, Shu M, Awadalla A, et al. XGEN-MM (blip-3): A family of open large multimodal models. ArXiv:2408.08872
8 Zhu S, Li S, Lei Y, et al. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. 13433–13447
9 Lan Z, Yu J, Li X, et al. Exploring better text image translation with multimodal codebook. ArXiv:2305.17415
10 Watanabe Y, Okada Y, Kim Y B, et al. Translation camera. In: Proceedings of the 14th International Conference on Pattern Recognition, 1998. 613–617
11 Yang J, Chen X, Zhang J, et al. Automatic detection and translation of text from natural scenes. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002
12 Afli H, Way A. Integrating optical character recognition and machine translation of historical documents. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), 2016. 109–116
13 Lan Z, Niu L, Meng F, et al. Translatotron-V(ison): An end-to-end model for in-image machine translation. In: Proceedings of Findings of the Association for Computational Linguistics ACL, 2024. 5472–5485
14 Mansimov E, Stern M, Chen M, et al. Towards end-to-end in-image neural machine translation. ArXiv:2010.10648
15 Li B, Zhu S, Wen L. MIT-10M: A large scale parallel corpus of multilingual image translation. In: Proceedings of the 31st International Conference on Computational Linguistics, 2025. 5154–5167
16 Jain P, Firat O, Ge Q, et al. Image translation network. Github.com, 2021
17 Ma C, Zhang Y, Tu M, et al. Improving end-to-end text image translation from the auxiliary text translation task. In: Proceedings of the 26th International Conference on Pattern Recognition (ICPR), 2022. 1664–1670
18 Ma C, Zhang Y, Tu M, et al. Multi-teacher knowledge distillation for end-to-end text image machine translation. In: Proceedings of International Conference on Document Analysis and Recognition, 2023. 484–501
19 Bai J, Bai S, Yang S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities. ArXiv:2308.12966
20 Chen Z, Wu J, Wang W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 24185–24198
21 Liu H, Li C, Li Y, et al. Llava-next: Improved reasoning, ocr, and world knowledge. 2024. https://llava-vl.github.io/blog/2024-01-30-llava-next/
22 Yao Y, Yu T, Zhang A, et al. Minicpm-v: A GPT-4v level mllm on your phone. ArXiv:2408.01800
23 Lu H, Liu W, Zhang B, et al. Deepseek-vl: Towards real-world vision-language understanding. ArXiv:2403.05525
24 Meta AI. Llama 3 model card. 2024. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
25 Bai S, Chen K, Liu X, et al. Qwen2. 5-vl technical report. ArXiv:2502.13923
26 Tang T, Luo W, Huang H, et al. Language-specific neurons: The key to multilingual capabilities in large language models. ArXiv:2402.16438
27 Goodfellow I J, Bulatov Y, Ibarz J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks. ArXiv:1312.6082
28 Zhang B, Xiong D, Su J, et al. Variational neural machine translation. ArXiv:1605.07869
29 Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation. ArXiv:1711.02281
30 OpenAI. Hello GPT-4O. 2024. https://openai.com/index/hellogpt-4o/
31 Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners. ArXiv:2109.01652
32 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 27730–27744
33 Xu L, Zhao Y, Zhou D, et al. Pllava: Parameter-free llava extension from images to videos for video dense captioning. ArXiv:2404.16994
34 Zhang C, Liu X, Jin M, et al. When AI meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. ArXiv:2407.18957
35 Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. In: Proceedings of ICLR, 2022
36 Yan L, Han C, Xu Z, et al. Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning. In: Proceedings of IJCAI, 2023. 1622–1630
37 Jie S, Wang H, Deng Z H. Revisiting the parameter efficiency of adapters from the perspective of precision redundancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 17217–17226
38 He X, Li C, Zhang P, et al. Parameter-efficient model adaptation for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 817–825
39 Liu S Y, Wang C Y, Yin H, et al. DoRA: Weight-decomposed low-rank adaptation. In: Proceedings of the 41st International Conference on Machine Learning, 2024. 32100–32121
40 French R. Catastrophic forgetting in connectionist networks. Trends Cogn Sci, 1999, 3: 128–135
41 Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA, 2017, 114: 3521–3526

42 Luo Y, Yang Z, Meng F, et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. ArXiv:2308.08747
43 Zhai Y, Tong S, Li X, et al. Investigating the catastrophic forgetting in multimodal large language models. ArXiv:2309.10313
44 Wang Z, Zhang Z, Ebrahimi S, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In: Proceedings of European Conference on Computer Vision, 2022. 631–648
45 Ju C, Han T, Zheng K, et al. Prompting visual-language models for efficient video understanding. In: Proceedings of European Conference on Computer Vision, 2022. 105–124
46 Han C, Wang Q, Cui Y, et al. E^2vpt: An effective and efficient approach for visual prompt tuning. ArXiv:2307.13770
47 Shen Y, Xu Z, Wang Q, et al. Multimodal instruction tuning with conditional mixture of LoRA. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 637–648
48 Wang T, Liu Y, Liang J C, et al. M^2PT: Multimodal prompt tuning for zero-shot instruction learning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024. 3723–3740
49 Molchanov P, Tyree S, Karras T, et al. Pruning convolutional neural networks for resource efficient inference. ArXiv:1611.06440
50 Xie W, Feng Y, Gu S, et al. Importance-based neuron allocation for multilingual neural machine translation. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2021
51 Cettolo M, Niehues J, Stüker S, et al. Report on the 11th IWSTL Evaluation Campaign. In: Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign, 2014. 2–17
52 Cettolo M, Federico M, Bentivogli L, et al. Overview of the IWSLT 2017 evaluation campaign. In: Proceedings of the 14th International Workshop on Spoken Language Translation, 2017. 2–14
53 Li C, Liu W, Guo R, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. ArXiv:2206.03001
54 NLLB Team, Costa-jussá M R, Cross J, et al. No language left behind: Scaling human-centered machine translation. Github.com, 2022
55 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 24824–24837
56 Niu L, Meng F, Zhou J. UMTIT: Unifying recognition, translation, and generation for multimodal text image translation. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024. 16953–16972
57 Liang Y, Zhang Y, Ma C, et al. Document image machine translation with dynamic multi-pre-trained models assembling. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024. 7084–7095
58 Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002. 311–318
59 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005. 65–72
60 Zheng Y, Zhang R, Zhang J, et al. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 400–410