LLaVA-NeuMT: A Layer-Aware Neuron Modulation Framework for Multimodal Multilingual Machine Translation

Anonymous ACL submission

Abstract

Multimodal Machine Translation (MMT) enhances translation quality by incorporating visual context, helping to resolve textual ambiguities. While existing MMT methods perform well in bilingual settings, extending them to multilingual translation remains challenging due to cross-lingual interference and ineffective parameter-sharing strategies. To address this, we propose LLaVA-NeuMT, a novel multimodal multilingual translation framework that explicitly models language-specific and language-agnostic representations to mitigate multilingual interference. Our approach consists of a layer selection mechanism that identifies the most informative layers for different language pairs and a neuron-level adaptation strategy that dynamically selects language-specific and agnostic neurons to improve translation quality while reducing redundancy. We conduct extensive experiments on the M3-Multi30K and M3-AmbigCaps datasets, demonstrating that LLaVA-NeuMT, while fine-tuning only 40% of the model parameters, surpasses full fine-tuning approaches and ultimately achieves SOTA results on both datasets. Our analysis further provides insights into the importance of selected layers and neurons in multimodal multilingual adaptation, offering an efficient and scalable solution to cross-lingual adaptation in multimodal translation.

1 Introduction

Machine translation has become increasingly crucial in our interconnected world, yet achieving accurate translations remains challenging due to the inherent ambiguities in natural language (Dabre et al., 2020; Klouchek and Batista-Navarro, 2024). A single word or phrase often carries multiple potential meanings, making it difficult for translation systems to select the appropriate interpretation without additional context. Multimodal Machine Translation (MMT) addresses this challenge by incorporating visual information alongside textual input, helping to resolve ambiguities and improve translation accuracy (Chen et al., 2021; Ma et al., 2022; Tayir et al., 2024). For example, as shown in Figure 1 (a), when translating the English sentence "There is a small house beside the bank" into German, purely text-based systems often misinterpret "bank" as a financial institution, producing "Neben der Bank steht ein kleines Haus." However, with access to the corresponding image, the system correctly recognizes "bank" as a riverbank and generates the accurate translation "Es gibt ein kleines Haus neben dem Ufer."

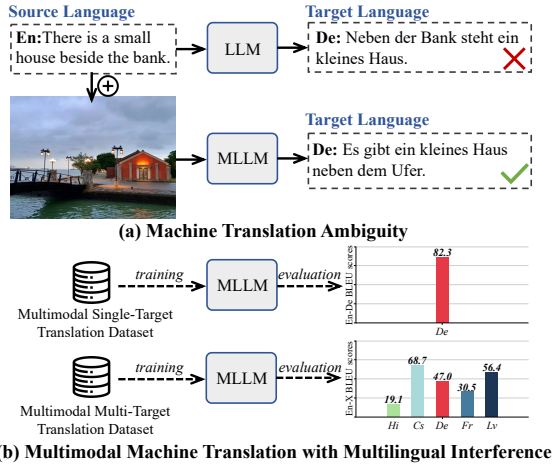

Figure 1: Challenges in MMT.

While MMT has demonstrated promising results in bilingual settings through various techniques such as multi-task learning, knowledge distillation, and attention mechanisms, extending these approaches to multilingual scenarios presents significant challenges(Fan et al., 2021; Wang et al., 2024c). Multilingual Neural Machine Translation (MNMT) has made progress in text-only translation by leveraging cross-lingual parameter sharing, evolving from simple parameter sharing to more sophisticated approaches like adaptive scheduling and language-specific modules(Jean et al., 2019; Pan et al., 2021; Feng et al., 2023). Recently, Mixture-

of-Experts (MoE) models have attempted to dynamically allocate computational resources across languages, but often struggle with overfitting and inefficient parameter utilization(Fedus et al., 2022; Li et al., 2023). Despite these advances, existing MNMT methods exclusively focus on text-based translation and do not address the unique complexities introduced by multimodal information.

As illustrated in Figure 1 (b), multimodal translation in multilingual settings introduces additional challenges beyond those found in either bilingual MMT or text-only MNMT. Recent studies have highlighted that indiscriminate parameter sharing in MNMT can lead to interference between languages, where high-resource languages dominate and degrade the performance of low-resource languages (Shaham et al., 2023; Li et al., 2023; Chen et al., 2024a). Furthermore, empirical analysis reveals that different layers in neural translation models serve distinct functions - lower layers often capture general linguistic patterns shared across languages, while higher layers learn language-specific and task-specific features (Tan et al., 2024; Zhu et al., 2024b). This layered hierarchy becomes particularly crucial in multilingual settings, as different language pairs may rely more heavily on certain layers for effective translation. However, current approaches treat all layers equally when sharing parameters across languages (Ma et al., 2023b; Lan et al., 2023; Tian et al., 2023), potentially leading to sub-optimal use of model capacity and increased interference. These observations raise critical questions: How can we identify and leverage the most relevant layers for each language pair? How should we balance parameter sharing across different layers to minimize interference while maintaining translation quality?

To address these challenges, we propose LLaVA-NeuMT, a framework designed to systematically identify and optimize the most relevant model components for each language pair. Our key insight is that selective parameter sharing at both the layer and neuron levels is crucial for balancing effective knowledge transfer and interference mitigation. Instead of sharing all parameters across languages indiscriminately, our method selectively determines which parts of the model are critical for each language pair. First, we introduce a layer selection mechanism that identifies the most informative layers for different language pairs, allowing the model to retain essential representations while reducing computational redundancy. Second, we propose a neuron-level adaptation strategy, where neurons within the selected layers are categorized as either language-specific or language-agnostic based on their activation and gradient variance. Finally, we design a training framework that selectively updates neurons based on the input language pair, mitigating inter-language interference while maintaining computational efficiency.

To validate our approach, we conduct extensive experiments on the M3-Multi30K (Guo et al., 2022) and M3-AmbigCaps (Li et al., 2021) datasets. The results show that LLaVA-NeuMT, utilizing only 40% of the model parameters, surpasses full fine-tuning baselines. By selecting key layers and fine-tuning language-specific and agnostic neurons, our approach achieves more effective multilingual adaptation. Furthermore, we visualize the importance of selected layers and neurons across languages, offering insights into the adaptation of multimodal translation models.

Our key contributions are as follows:

- We propose **LLaVA-NeuMT**, a multimodal multilingual translation framework that explicitly models *language-specific* and *language-agnostic* representations to mitigate cross-lingual interference in multimodal translation.
- We introduce a **layer and neuron selection mechanism** that identifies the most informative layers and neurons for each language pair, effectively preserving critical representations while reducing redundancy.
- We achieve **SOTA translation performance** across multiple language pairs while fine-tuning a subset of model parameters. Additionally, our analysis provides insights into the importance of different layers and neurons in multimodal multilingual adaptation.

2 Related Work

Multimodal Machine Translation Multimodal Machine Translation (MMT) enhances translation quality by integrating visual context to resolve linguistic ambiguities. Prior research has explored four primary approaches: multi-task learning, knowledge distillation, contrastive learning, and attention-based mechanisms. Multi-task learning integrates OCR and translation models to improve cross-modal representation learning, but these methods often struggle with efficient multilingual adaptation (Chen et al., 2021; Ma et al.,

2022; Su et al., 2021). To address this, adaptive mechanisms have been introduced to bridge modality gaps and enhance translation consistency (Ma et al., 2023b; Lan et al., 2023). Knowledge distillation has been widely used to transfer multimodal knowledge from teacher to student models, ensuring better generalization but often increasing computational overhead (Chen et al., 2023; Ma et al., 2023c). Contrastive learning further refines OCR-text alignment and improves robustness in translation tasks, yet remains constrained by reliance on predefined feature mappings (Ma et al., 2024; Peng et al., 2022). Attention-based mechanisms dynamically focus on relevant image regions, improving semantic grounding, but they lack efficient parameter selection for multilingual translation (Mansimov et al., 2020; Hinami et al., 2021; Jain et al., 2021; Tian et al., 2023). While these methods enhance machine translation performance, they often overlook computational efficiency in large-scale multimodal models. As computational demands grow with model size and multilingual adaptation, recent works have emphasized the need to balance model capacity with efficiency (Liu et al., 2022; Ma et al., 2023a). However, existing approaches still lack fine-grained control over language-specific and agnostic parameters. To address these challenges, we propose a layer-aware neuron modulation framework that improves translation efficiency while optimizing parameter utilization.

Multilingual Neural Machine Translation Multilingual Neural Machine Translation (MNMT) enables translation across multiple languages within a single model but faces challenges such as inter-language interference and capacity bottlenecks (Aharoni et al., 2019; Fan et al., 2021; Wei et al., 2024). Prior works address these issues through adaptive scheduling (Jean et al., 2019; Pan et al., 2021), gradient-based optimization (Wang et al., 2020; Feng et al., 2023), and language-specific modules (Philip et al., 2020; Zhang et al., 2021). Mixture-of-Experts (MoE) models allocate capacity dynamically (Fedus et al., 2022; Li et al., 2023), though overfitting remains a concern. Recent studies highlight that indiscriminate parameter sharing degrades high-resource language performance (Huang et al., 2024; Nimma et al., 2024; Javed et al., 2025), leading to strategies such as binary masks (Poppi et al., 2024) and contrastive learning (Liang et al., 2024) to mitigate interference. However, these approaches often introduce additional complexity and computational costs. While research on multilingual interference has primarily focused on text-based models (Jean et al., 2019; Li et al., 2023; Javed et al., 2025), its implications for multimodal translation remain insufficiently studied. In contrast, we introduce a selective layer and neuron-level modulation framework to optimize multilingual adaptation, reducing interference while maintaining efficiency in multimodal MNMT.

3 Methodology

3.1 Multimodal Machine Translation

Multimodal machine translation extends traditional machine translation by incorporating visual information to enhance contextual understanding. Given a source sentence X^s in language s, a corresponding image I, and a target language t, the objective is to generate a translated sentence Y^t that preserves the semantics of the source sentence while leveraging visual context. The translation process can be formulated as a function $\mathcal{F}$ that maps the source text and image to the target text:

$$Y^t = \mathcal{F}(X^s, I, s, t; \theta), \tag{1}$$

where θ represents the model parameters. The model encodes textual features through a text encoder $\mathcal{E}_t$ and extracts visual features using a vision encoder $\mathcal{E}_v$:

$$\mathcal{T} = \mathcal{E}_t(X^s), \quad \mathcal{V} = \mathcal{E}_v(I). \tag{2}$$

The extracted textual and visual features are combined within a multimodal translation model, producing an intermediate representation that is subsequently decoded into the target language.

3.2 Selecting Effective Layers of the Model

In MMT, different layers of the model contribute differently to text and image processing. To improve efficiency while maintaining translation quality, and inspired by (Li et al., 2024; Wang et al., 2024b), we introduce a layer selection method that identifies and retains the most informative layers in both the vision-language connector and the large language model (LLM). Given a model with L layers, the objective is to determine a subset $\mathcal{L} \subseteq \{1, 2, \ldots, L\}$ that maximizes task relevance while reducing redundancy.

The importance of each layer is assessed based on activation similarity before and after supervised fine-tuning (SFT), as illustrated in Figure 2 (a).

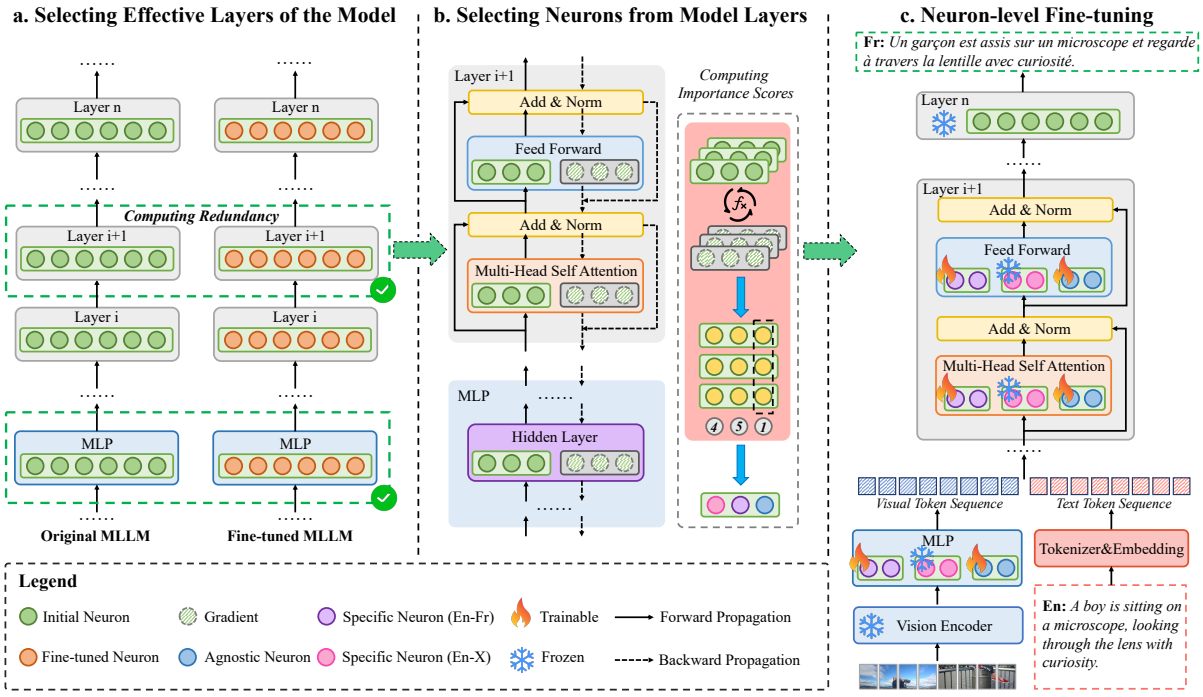

Figure 2: LLaVA-NeuMT Model Architecture.

Given activations from layer l in the pretrained model (X_l^A) and the fine-tuned model (X_l^B), the redundancy-based importance score R_l is computed as:

$$R_l = \frac{(X_l^A \cdot X_l^B)^2}{\|X_l^A\|^2 \|X_l^B\|^2 + \epsilon}, \tag{3}$$

where X_l^A and X_l^B are the activations from layer l in the pretrained and fine-tuned models, and ϵ is a small constant to avoid numerical instability. A lower R_l value suggests that a layer undergoes significant adaptation during fine-tuning, indicating its importance for the translation task. Layers are ranked based on R_l, and a subset $\mathcal{L}_s$ is selected corresponding to the top α fraction of layers, where α is a tunable hyperparameter.

3.3 Selecting Important Neurons of the Model

In MMT, different neurons within each selected layer contribute differently to various language pairs (Zhu et al., 2024a). To further optimize the model, we introduce a neuron selection mechanism that identifies the most relevant neurons for each language pair while preserving generalizable neurons across all languages. Given a layer l with N neurons, our objective is to classify neurons into two categories: *language-specific neurons* and *language-agnostic neurons*.

Neuron selection is performed on the previously selected layers, as illustrated in Figure 2 (b). The importance of each neuron n is evaluated based on its activation and gradient values during fine-tuning. Specifically, for each training instance, we compute the importance score as:

$$\mathcal{I}_n = |A_n \times G_n| \tag{4}$$

where A_n and G_n represent the activation and backpropagation gradient of neuron n, respectively. Given K language pairs, we aggregate the importance scores over T training samples and compute the variance across languages as:

$$\sigma^2(n) = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{I}_n^k - \bar{\mathcal{I}}_n \right)^2 \tag{5}$$

where $\mathcal{I}_n^k$ denotes the importance score of neuron n for language pair k, and $\bar{\mathcal{I}}_n$ represents the mean importance score across all language pairs.

To classify neurons, we first define $\mathcal{N}$ as the set of all neurons in the selected layers:

$$\mathcal{N} = \{n \mid n \in \mathcal{N}_l, l \in \mathcal{L}_s\} \tag{6}$$

where $\mathcal{L}_s$ is the selected layer set and $\mathcal{N}_l$ the neurons in layer l. We then define two subsets:

$$\mathcal{S}_k = \{n \in \mathcal{N} \mid \mathcal{I}_n^k = \max_j \mathcal{I}_n^j\} \tag{7}$$

where $\mathcal{S}_k$ represents the set of language-specific neurons, which exhibit the highest importance for a single language pair k.

$$\mathcal{A} = \{n \in \mathcal{N} \mid \sigma^2(n) \leq \epsilon\} \tag{8}$$

where $\mathcal{A}$ represents the set of language-agnostic neurons, which maintain relatively stable importance scores across all language pairs, with a variance threshold ϵ.

3.4 LLaVA-NeuMT

LLaVA-NeuMT performs neuron-level adaptation based on the previously selected layers and classified neurons, as illustrated in Figure 2 (c). Given a source sentence X^s in language s, an image I, and a target language t, the model extracts features using a text encoder $\mathcal{E}_t$ and a vision encoder $\mathcal{E}_v$, producing textual and visual representations as defined in Equation (2). These representations are passed through the selected layers $\mathcal{L}_s$ (Section 3.2), where neuron updates are applied selectively. Based on the classification of neurons into language-specific ($\mathcal{S}_k$) and language-agnostic ($\mathcal{A}$) categories (Section 3.3), only relevant neurons receive parameter updates during fine-tuning.

To achieve this, a gradient masking mechanism is applied to constrain updates to neurons belonging to $\mathcal{S}_k$ or $\mathcal{A}$. Specifically, for each neuron n, the gradient is modified as follows:

$$G'_n = \begin{cases} G_n, & \text{if } n \in \mathcal{A} \cup \mathcal{S}_k \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where G_n represents the computed gradient of neuron n. Neurons outside these sets are frozen, preventing unnecessary parameter updates.

The final model update is performed using:

$$\theta_n \leftarrow \theta_n - \eta G'_n \tag{10}$$

where η is the learning rate. By restricting updates to selected neurons, LLaVA-NeuMT efficiently adapts the model while maintaining stability across different language pairs. This fine-tuning strategy ensures that multimodal representations are effectively adapted, allowing both textual and visual features to be optimized for MNMT.

4 Experiments

4.1 Experimental Setting

Datasets We evaluate our approach on two multimodal MNMT datasets: M3-Multi30K (Guo et al., 2022) and M3-AmbigCaps (Li et al., 2021). M3-Multi30K consists of 29,000 image-text translation pairs for training and 1,000 for testing, covering multiple language pairs. M3-AmbigCaps is a larger dataset with 89,600 training pairs and 1,000 test pairs, designed for evaluating multimodal translation performance.

Experimental Setup We adopt LLaVA-1.5-7B (Liu et al., 2024) as the pretrained backbone and optimize training using DeepSpeed ZeRO-3 on 4 × A100 (80GB) GPUs. The model is trained for 4 epochs with a per-device batch size of 16 and a gradient accumulation step of 1. Mixed precision training with BF16 is applied to reduce memory overhead. The optimizer is AdamW with a learning rate of 2e-5 and a cosine annealing scheduler, with 3% warmup. Weight decay is set to 0. The maximum text sequence length is 2048, and image inputs are resized to a fixed aspect ratio, with visual features extracted from the second-to-last layer of the Vision Transformer (ViT) (Nguyen et al., 2024).

Evaluation Metrics & Baselines We evaluate translation performance using BLEU-4. Baselines include Text-only MT models: Text Transformer (Fan et al., 2021); Open-source MMT models: Qwen2-VL (Wang et al., 2024a), MiniCPM (Yao et al., 2024), InternVL (Chen et al., 2024b); Closed-source MMT models: GPT-4o (Achiam et al., 2023), Gemini-1.5-Pro (Team et al., 2024); and Multimodal MNMT models: Vision Matters (Gated Fusion) (Li et al., 2021), Vision Matters (Concatenation) (Li et al., 2021), LVP-M3 (Guo et al., 2022), and the multilingual fine-tuned version of LLaVA-1.5 (Liu et al., 2024).

4.2 Main Results and Analysis

We evaluate the performance of different models across four categories, as shown in Table 1 and Table 2, Text-only MT achieves strong results, demonstrating that textual models alone can provide high-quality translations. However, it still underperforms compared to Multimodal MNMT, which integrates visual context to improve translation quality. Open-source MMT models show significantly lower performance, particularly in low-resource languages such as Latvian, Hindi, and Turkish, likely due to the lack of multimodal multilingual training data, which limits their generalization in multilingual settings. Closed-source MMT models, such as GPT-4o, achieve competitive results in high-resource languages but show a noticeable drop in low-resource scenarios, suggesting that general-purpose multimodal models are not optimized for multilingual translation. In contrast, Mul-

Type	Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg-all
Text-only MT	Text Transformer (Fan et al., 2021)	61.8	32.8	40.6	51.2	59.0	53.8	49.8
Open-source MMT	Qwen2-VL-7B (Wang et al., 2024a)	44.1	7.8	33.5	0.1	0.6	0.8	14.5
	MiniCPM-2.6-8b (Yao et al., 2024)	26.2	4.0	27.2	0.1	0.2	0.3	9.7
	InternVL-2.5-7b (Chen et al., 2024b)	35.2	8.2	25.8	0.1	3.3	1.0	12.3
Closed-source MMT	GPT-4o (Achiam et al., 2023)	53.8	**37.4**	**44.3**	39.4	28.3	28.6	38.6
	Gemini-1.5-Pro (Team et al., 2024)	38.5	22.2	23.5	24.3	10.4	22.4	23.6
Multimodal MNMT	Vision Matters (Gated fusion) (Li et al., 2021)	62.5	32.9	41.2	52.1	59.6	54.2	50.4
	Vision Matters (Concatenation) (Li et al., 2021)	59.7	33.1	39.8	50.3	57.6	51.4	48.6
	LVP-M3 (Guo et al., 2022)	63.7	34.6	43.2	53.5	61.4	55.6	52.0
	LLaVA-1.5-SFT(default) (Liu et al., 2024)	66.5	35.9	42.2	56.1	61.5	57.8	53.3
Ours	LLaVA-NeuMT (40%)	**67.0**	36.0	42.0	57.3	60.0	58.3	53.4
	LLaVA-NeuMT (80%)	66.8	35.9	42.6	**58.2**	**61.8**	**60.7**	**54.3**

Table 1: BLEU scores on the M3-Multi30K test set. Best results are in **bold**, second-best are underlined.

Type	Model (En→X)	Fr	Cs	De	Lv	Hi	Tr	Avg_all
Text-only MT	Text Transformer (Fan et al., 2021)	62.3	47.8	49.0	46.6	52.4	35.9	49.0
Open-source MMT	Qwen2-VL-7B (Wang et al., 2024a)	40.3	2.7	27.3	0.3	0.6	0.7	12.0
	MiniCPM-2.6-8b (Yao et al., 2024)	32.6	2.8	19.8	0.1	0.15	0.2	9.3
	InternVL-2.5-7b (Chen et al., 2024b)	31.6	6.16	10.7	0.1	3.3	0.8	8.8
Closed-source MMT	GPT-4o (Achiam et al., 2023)	43.6	29.6	38.0	26.5	24.9	16.7	29.9
	Gemini-1.5-Pro (Team et al., 2024)	28.8	13.6	18.3	15.9	12.2	12.2	16.8
Multimodal MNMT	Vision Matters (Gated fusion) (Li et al., 2021)	64.3	50.3	51.2	48.5	54.1	38.7	51.2
	Vision Matters (Concatenation) (Li et al., 2021)	62.6	47.6	48.7	45.9	52.7	36.0	48.9
	LVP-M3 (Guo et al., 2022)	65.7	52.9	53.7	51.6	56.3	42.7	53.8
	LLaVA-1.5-SFT(default) (Liu et al., 2024)	72.1	57.3	60.3	56.5	56.8	45.2	58.0
Ours	LLaVA-NeuMT (40%)	73.2	57.0	60.9	56.2	56.5	46.2	58.3
	LLaVA-NeuMT (80%)	**74.1**	**58.4**	**61.7**	**57.8**	**58.4**	**47.9**	**59.7**

Table 2: BLEU scores on the M3-AmbigCaps test set. Best results are in **bold**, second-best are underlined.

timodal MNMT models consistently achieve better BLEU scores, confirming that incorporating multimodal signals benefits multilingual translation. Among them, LLaVA-1.5-SFT enhances translation quality through supervised fine-tuning. Our proposed LLaVA-NeuMT further improves performance while fine-tuning only 40% of the model parameters, demonstrating the efficiency of selective layer adaptation. When increasing the fine-tuned layers to 80%, LLaVA-NeuMT achieves the best results, showing that balancing layer selection and neuron modulation enhances translation performance while maintaining efficiency. Additionally, our fine-tuning strategy, which adjusts language-specific and agnostic neurons at a 1:9 ratio, ensures effective multilingual adaptation. In terms of language-specific trends, GPT-4o performs well on high-resource languages such as French, Czech, and German in the M3-Multi30K test set but struggles in lower-resource languages. The performance gap is more evident in the M3-AmbigCaps test set, where the larger dataset scale and increased task complexity further challenge general-purpose models. By contrast, LLaVA-NeuMT consistently outperforms other models across both datasets, demonstrating its robustness in Multimodal MNMT.

4.3 Effect of Layer Selection on MMT

To investigate the role of layer selection in multimodal multilingual translation, we evaluate performance by selecting the top 20%, 40%, 60%, 80%, and 100% most important layers, ranked by importance scores computed in Section 3.2. In this experiment, the neuron selection strategy remains fixed, with language-specific and agnostic neurons adjusted at a 1:9 ratio, ensuring that the only variable is the number of selected layers. As shown in Figure 3, BLEU scores increase as more layers are included, reaching the highest performance at 80% selection. Beyond this point, performance declines, suggesting that retaining all layers introduces redundancy or noise, negatively impacting transla-

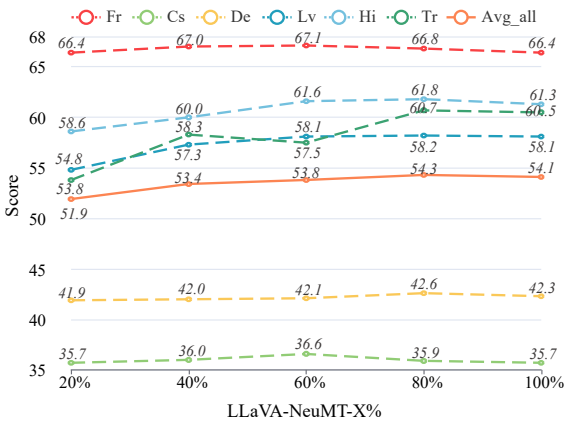

Figure 3: Effect of layer selection on translation. x% indicates the top x% most important layers.

En→X	Fr	Cs	De	Lv	Hi	Tr	Avg_all
Standard	66.4	35.7	42.3	58.1	61.3	60.5	54.1
Agnostic	65.8	35.6	41.2	56.3	61.1	59.7	53.3
Specific	65.9	35.2	42.0	57.0	60.5	61.3	53.7

Table 3: Effect of agnostic and specific neurons on multimodal multilingual translation on the M3-Multi30K dataset. "Standard" denotes a 1:9 specific-to-agnostic neuron ratio, while "Agnostic" and "Specific" refer to models fine-tuning only agnostic or specific neurons.

tion quality. Using only 20% of the layers leads to significantly lower BLEU scores, indicating that a minimal subset is insufficient for effective multimodal multilingual adaptation. Between 40% and 80%, all language pairs exhibit consistent improvements, with the most pronounced gains observed in low-resource languages such as Latvian, Hindi, and Turkish. For high-resource languages such as French, Czech, and German, performance stabilizes beyond 60% and slightly decreases at 100%, reinforcing that excessive layers do not necessarily contribute positively to translation. These findings demonstrate that an optimal layer selection strategy enhances translation quality while maintaining efficiency, with 80% selection striking the best balance between performance and computational cost.

4.4 Effect of Neuron Selection on MMT

To analyze the impact of agnostic and specific neurons in multimodal multilingual translation, we conduct experiments where all layers are selected while varying the neurons that are fine-tuned. As shown in Table 3, the highest BLEU score is achieved when both neuron types are optimized in a 1:9 ratio. Fine-tuning only agnostic neurons results in a slight performance drop, while fine-tuning only specific neurons leads to a further decline. This suggests that while specific neurons contribute to translation quality, agnostic neurons play a more crucial role in ensuring multilingual adaptation. The relatively competitive performance of fine-tuning specific neurons alone indicates that language-specific features remain valuable, particularly in distinguishing linguistic variations. However, the performance gap between agnostic-only and specific-only settings reinforces the greater importance of agnostic neurons in maintaining stable multilingual translation. Examining language-specific trends, Czech benefits more from fine-tuning agnostic neurons, suggesting a stronger dependence on cross-lingual representations, whereas Turkish achieves its highest accuracy when only specific neurons are fine-tuned, indicating that some languages rely more on task-specific adaptation.

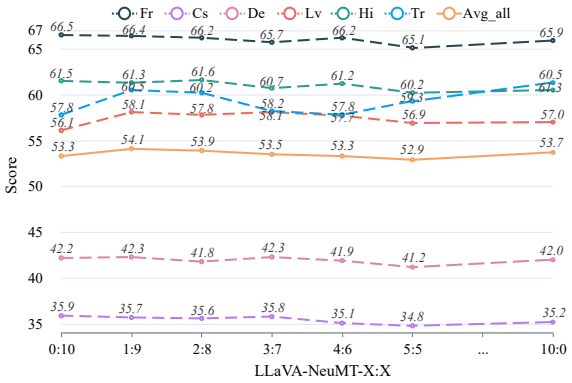

Figure 4: Impact of specific-to-agnostic neuron ratio on translation performance on the M3-Multi30K dataset.

To further examine the effect of adjusting the ratio of specific to agnostic neurons, we conduct experiments while keeping all layers selected. As shown in Figure 4, increasing the proportion of specific neurons initially improves BLEU scores, peaking at a 1:9 ratio. Beyond this point, performance declines as the proportion of agnostic neurons decreases, suggesting that excessive specific neurons may reduce generalization ability. However, at extreme ratios (e.g., 10:0), performance slightly rebounds, indicating that in certain cases, heavily relying on specific neurons can still capture relevant translation patterns. This suggests that while an optimal balance of neuron types is necessary, models exhibit some degree of robustness when specific neurons dominate. Across different language pairs, Czech exhibits a steady decline when agnostic neurons are reduced, confirming its reliance on agnostic representations. In contrast, Hindi and Turkish maintain relatively stable performance across different neuron ratios, demonstrating adaptability to

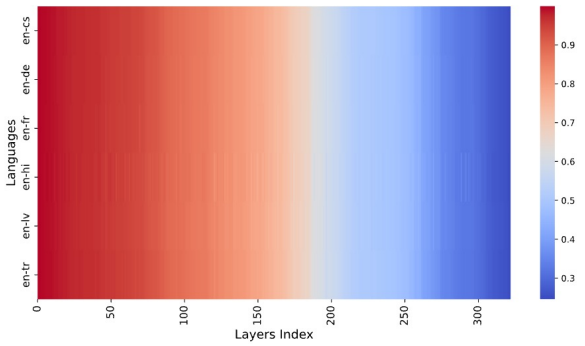

Figure 5: Layer importance visualization in multimodal multilingual translation on the M3-Multi30K dataset.

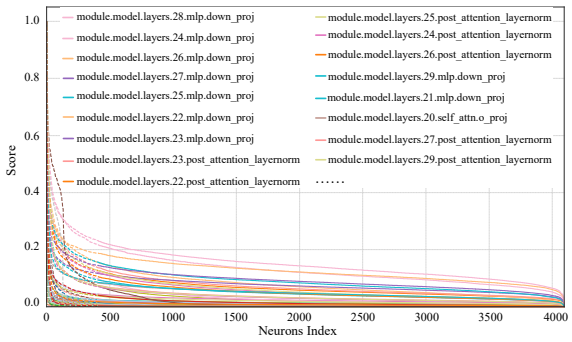

Figure 6: Neuron importance in multimodal multilingual translation on M3-Multi30K dataset.

both neuron types. These findings emphasize the necessity of a well-balanced allocation of agnostic and specific neurons for optimal MMT.

4.5 Visualization of Layer Importance

To investigate the role of different layers in multimodal multilingual translation, we visualize the layer importance scores across multiple language pairs in Figure 5. The x-axis represents model layers, the y-axis denotes language pairs, and the color intensity indicates relative importance. The heatmap reveals significant variations in layer importance across the model, demonstrating that selecting key layers is necessary rather than uniformly fine-tuning all layers. From a horizontal perspective, the first 250 layers (approximately 80% of the model depth) exhibit relatively high importance scores, consistently exceeding 0.5. This trend aligns with the findings in Section 4.3, where selecting the top 40-80% of layers resulted in optimal translation performance. The concentration of importance in these layers suggests that they capture essential multimodal and multilingual representations. From a vertical perspective, the importance scores remain relatively stable across different language pairs, indicating that layer selection is primarily influenced by architectural properties rather than specific language characteristics. This confirms that an effective layer selection can enhance computational efficiency without significantly affecting translation quality across languages.

4.6 Analysis of Specific and agnostic neurons

To investigate the distinction between specific and agnostic neurons in multimodal multilingual translation, we visualize neuron importance variance across six language pairs in Figure 6. We select the top 40% of layers (108 layers) and observe that in each layer, a small subset of neurons exhibits significantly higher variance, indicating their language specificity. This confirms the necessity of differentiating specific and agnostic neurons rather than treating them uniformly. From a distribution perspective, the first 10% of neurons in each layer (dashed lines) display high variance, while the remaining 90% (solid lines) maintain stable scores. This supports our choice of a 1:9 ratio between specific and agnostic neurons, ensuring an optimal balance between language adaptability and cross-lingual generalization. Furthermore, we identify key neuron types critical to multimodal multilingual translation, including attention projection layers and MLP down-projection layers. Unlike conventional large language models, which primarily rely on deep linguistic representations, multimodal translation models emphasize connector layers for effective cross-modal alignment, underscoring their importance in improving translation quality.

5 Conclusion

In this work, we tackled multilingual interference in MMT by introducing LLaVA-NeuMT, a framework that selectively optimizes layers and neurons to enhance efficiency and translation quality. Our approach integrates a layer selection mechanism to retain the most informative layers and a neuron-level adaptation strategy to balance language-specific and agnostic representations. Experiments on the M3-Multi30K and M3-AmbigCaps datasets show that LLaVA-NeuMT achieves SOTA performance while fine-tuning fewer parameters. Further analysis reveals that selecting 40-80% of layers yields optimal results, and a 1:9 specific-to-agnostic neuron ratio effectively balances generalization and adaptation. Future work will explore adaptive parameter-sharing strategies and extend our approach to broader multilingual and multimodal scenarios.

Limitations

While our proposed LLaVA-NeuMT framework has demonstrated strong performance in multimodal MNMT, several aspects remain worth exploring. Our current approach selects layers and neurons based on fixed thresholds, such as choosing the top 40% of layers and applying a predefined ratio of specific to agnostic neurons. While effective, this static strategy may not be optimal across different language pairs and translation contexts. Future work could explore more adaptive selection mechanisms to further enhance efficiency and generalization. Additionally, beyond multilingual settings, our approach could be extended to balance general-purpose language tasks with domain-specific translation challenges in multimodal scenarios, addressing broader applications of multimodal translation.

References

Josh Achiam, Steven Adler, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL*, pages 3874–3884.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mclip: Multilingual clip via cross-lingual transfer. In *ACL*, pages 13028–13043.

Liang Chen, Shuming Ma, Dongdong Zhang, et al. 2024a. On the pareto front of multilingual neural machine translation. *NeurIPS*, 36.

Zhe Chen, Jiannan Wu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198.

Zhuo Chen, Fei Yin, Xu-Yao Zhang, Qing Yang, and Chena-Lin Liu. 2021. Cross-lingual text image recognition via multi-task sequence to sequence learning. In *ICPR*, pages 3122–3129.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys*, 53(5):1–38.

Angela Fan, Shruti Bhosale, Holger Schwenk, et al. 2021. Beyond english-centric multilingual machine translation. *JMLR*, 22(107):1–48.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39.

Xiaocheng Feng, Xinwei Geng, Baohang Li, Bing Qin, et al. 2023. Towards higher pareto frontier in multilingual machine translation. In *ACL*.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, et al. 2022. Lvp-m3: Language-aware visual prompt for multilingual multimodal machine translation. In *EMNLP*, pages 2862–2872.

Ryota Hinami, Shonosuke Ishiwatari, et al. 2021. Towards fully automated manga translation. In *AAAI*, volume 35, pages 12998–13008.

Kaiyu Huang, Fengran Mo, Hongliang Li, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. Image translation network.

Arifa Javed, Hongying Zan, Orken Mamyrbayev, Muhammad Abdullah, et al. 2025. Transformer-based re-ranking model for enhancing contextual and syntactic translation in low-resource neural machine translation. *Electronics*, 14(2):243.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. *arXiv preprint arXiv:1909.06434*.

Bozhidar Klouchek and Riza Theresa Batista-Navarro. 2024. Bulgarian grammar error correction with data augmentation and machine translation techniques. In *ICNLSP 2024*, pages 365–376.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *ACL*, pages 3479–3491.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *EMNLP*, pages 8556–8562.

Shangjie Li, Xiangpeng Wei, Shaolin Zhu, et al. 2023. Mmnmt: Modularizing multilingual neural machine translation with flexibly assembled moe and dense blocks. In *EMNLP*, pages 4978–4990.

Wei Li, Lujun Li, Mark G Lee, and Shengjie Sun. 2024. Adaptive layer sparsity for large language models via activation correlation assessment. In *NeurIPS*.

Yunlong Liang, Fandong Meng, Jiaan Wang, et al. 2024. Continual learning with semi-supervised contrastive distillation for incremental neural machine translation. In *ACL*, pages 10914–10928.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Jiaheng Liu, Haoyu Qin, Yichao Wu, Jinyang Guo, et al. 2022. Coupleface: Relation matters for face recognition distillation. In *ECCV*, pages 683–700.

Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2024. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task. In *ICPR*, pages 1664–1670.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023a. Ccim: Cross-modal cross-lingual interactive image translation. In *EMNLP 2023*, pages 4959–4965.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. E2timt: Efficient and effective modal adapter for text image machine translation. In *ICDAR*, pages 70–88.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023c. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *ICDAR*, pages 484–501.

Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *NLPBT*, pages 70–74.

Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R Oswald, Cees GM Snoek, and Xinlei Chen. 2024. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. *arXiv preprint arXiv:2406.09415*.

Divya Nimma, VK Somasekhar Srinivas, Santoshi Sen Gupta, Hari Nair, R Latha Devi, and B Kiran Bala. 2024. Comparative analysis of deep learning models for multilingual language translation. In *SLAAI-ICAI*, pages 1–6.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *ACL*, pages 244–258.

Ru Peng, Yawen Zeng, and Jake Zhao. 2022. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation. In *EMNLP*, pages 2379–2390.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *EMNLP*, pages 4465–4470.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2024. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *arXiv preprint arXiv:2410.18210*.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *ACL*, pages 15849–15863.

Tonghua Su, Shuchen Liu, and Shengjie Zhou. 2021. Rtnet: An end-to-end method for handwritten text image translation. In *ICDAR*, pages 99–113.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Transactions on Artificial Intelligence*.

Gemini Team, Petko Georgiev, , et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. In-image neural machine translation with segmented pixel sequence-to-sequence model. In *EMNLP*, pages 15046–15057.

Peng Wang, Shuai Bai, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qiufeng Wang, Xu Yang, Fu Feng, Xin Geng, et al. 2024b. Cluster-learngene: Inheriting adaptive clusters for vision transformers. In *NeurIPS*.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *ACL*, pages 8526–8537.

Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. 2024c. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *ACM MM*, pages 8296–8305.

Jingxuan Wei, Linzhuang Sun, Yichong Leng, Xu Tan, Bihui Yu, and Ruifeng Guo. 2024. Sentence-level or token-level? a comprehensive study on knowledge distillation. In *IJCAI*, pages 6531–6540.

Yuan Yao, Tianyu Yu, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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

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 *ACL*, pages 12135–12148.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, et al. 2024b. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.