Language-Specific Layer Matters: Efficient Multilingual Enhancement for Large Vision-Language Models

Yuchun Fan[1], Yilin Wang[1], Yongyu Mu[1], Lei Huang[4], Bei Li[3], Xiaocheng Feng[4], Tong Xiao[1,2]*, Jingbo Zhu[1,2]

[1] NLP Lab, School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
[3]Meituan Inc. [4]Harbin Institute of Technology, Harbin, China
yuchunfan_neu@outlook.com {xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Large vision-language models (LVLMs) have demonstrated exceptional capabilities in understanding visual information with human languages but also exhibit an imbalance in multilingual capabilities. In this work, we delve into the multilingual working pattern of LVLMs and identify a salient correlation between the multilingual understanding ability of LVLMs and language-specific neuron activations in shallow layers. Building on this insight, we introduce PLAST, a training recipe that achieves efficient multilingual enhancement for LVLMs by **P**recise **LA**nguage-**S**pecific layers fine-**T**uning. PLAST first identifies layers involved in multilingual understanding by monitoring language-specific neuron activations. These layers are then precisely fine-tuned with question-translation pairs to achieve multilingual alignment. Our empirical results on MMBench and MMMB demonstrate that PLAST effectively improves the multilingual capabilities of LVLMs and achieves significant efficiency with only 14% of the parameters tuned. Further analysis reveals that PLAST can be generalized to low-resource and complex visual reasoning tasks, facilitating the language-specific visual information engagement in shallow layers[1].

1 Introduction

Large vision-language models (LVLMs) have made remarkable progress in understanding visual information with human languages, achieving impressive performance in multimodal tasks like visual question answering (VQA) (Liu et al., 2024a,c; Lin et al., 2024). However, they still struggle in multilingual scenarios because the training corpora are predominantly English-centric (Chen et al., 2023a). This imbalance poses significant challenges for their real-world applications across the globe.

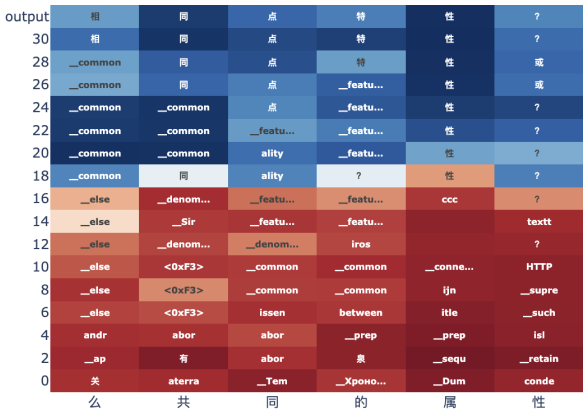


Figure 1: The logit lens (Nostalgebraist, 2020) applies the language modeling head (LLaVA-1.5-7B in our case) to intermediate layer embeddings, producing one next-token distribution per position (x-axis) and per layer (y-axis). The blue-to-red gradient indicates entropy levels, ranging from low to high. Full visualizations for all languages are shown in Figure 9–13.

Recent efforts to enhance the multilingual abilities of LVLMs primarily focus on two aspects. One line of research (Qin et al., 2023; Zhang et al., 2024b; Guo et al., 2023; Mu et al., 2023) involves incorporating translation, either implicitly or explicitly, into the design of prompts, encouraging the model to first comprehend questions in English and then solve them step by step. However, the inadequate multilingual understanding capability of LVLMs often triggers cascading errors, leading to inferior performance. Another approach expands this idea by adopting a *translate-then-train* strategy (Sun et al., 2025; Qu et al., 2024), where English training data are first translated into multiple languages via machine translation, followed by fine-tuning for multilingual instruction alignment. Despite effectiveness, these approaches rely on accurately translating large-scale complex multimodal data and full-parameter fine-tuning, which limits their applicability in data-deficiency and resource-constrained scenarios.

* Corresponding author.
[1]The project will be available at: https://github.com/fmm170/PLAST

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12473–12500
November 4-9, 2025

When considering multilingual models, one might think of capturing language-specific representation in certain parts of these models. Inspired by recent studies (Tang et al., 2024; Zhao et al., 2024; Zhao and Zhang, 2024) on the multilingual mechanisms in large language models (LLMs), which have verified that the multilingual working pattern of LLMs can be divided into three stages: language understanding, task-solving, and language converting. Moreover, during this process, the activation of language-specific neurons (Fan et al., 2025) can serve as a crucial indicator for distinguishing each working stage. We further investigate *whether this pattern persists in LVLMs* (§2). Our pilot study reveals that LVLMs also follow this three-stage process for handling multilingualism, with lower-level layers more engaged in learning language-specific representations. As shown in Figure 1, the Chinese query undergoes an explicit process of language understanding (Zh $\rightarrow$ En) and language converting (En $\rightarrow$ Zh). This motivates us to unlock the multilingual capabilities more efficiently by precisely enhancing language-specific representations.

To this end, we propose PLAST, a training recipe designed to achieve efficient multilingual enhancement for LVLMs, as outlined in Figure 2. Specifically, starting with a visual instruction-tuned model, we first identify decoder layers mostly involved in language understanding by monitoring the activation of language-specific neurons using multilingual image-text pairs (§3.1). More concretely, we calculate the mean squared deviation (MSD) of the numbers of neurons activated by different languages and select the layers with higher MSD scores. Subsequently, PLAST precisely fine-tunes only these identified layers using *a small amount of* image-question translation pairs (§3.2). This improves its multilingual understanding ability by precisely enhancing the language-specific representation while not affecting task-solving abilities at higher layers.

To evaluate the effectiveness of PLAST, we conduct extensive experiments on two multilingual VQA benchmarks, MMBench (Liu et al., 2024c) and MMMB (Sun et al., 2025) across three LVLMs. Experimental results demonstrate the effectiveness of PLAST in improving multilingual performance, with average gains of 8.0% and 4.0% on MMBench and MMMB. Furthermore, compared to full-parameter fine-tuning approaches, PLAST achieves superior efficiency with only 15.6% and 12.5% of the parameters within 7B and 13B models tuned. Further analysis reveals that PLAST can be generalized to low-resource and complex visual reasoning tasks, facilitating the language-specific visual information engagement in shallow layers.

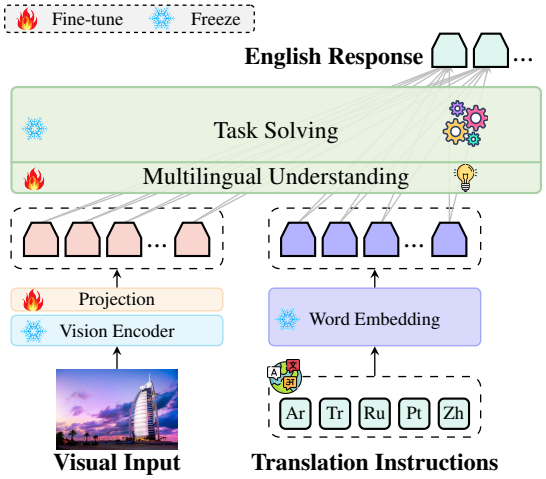


Figure 2: An overview of our method, PLAST. denotes the question-translation data. For instance, the instruction for training is: "Translate this from Chinese to English: 这座建筑是什么样子的?".

2 Preliminaries

Recent studies (Tang et al., 2024; Fan et al., 2025) have underscored the pivotal role of language-specific neurons within LLMs for multilingual processing. Drawing from these findings, this section starts with a preliminary analysis aimed at exploring whether neurons within LVLMs exhibit language-specific behavior and understanding their significance for multilingual visual understanding.

2.1 Notation

LVLMs, represented by the LLaVA series (Liu et al., 2023), typically comprise a vision encoder with a pre-trained LLM. Given a visual input $\boldsymbol{X}_v$ and a textual input $\boldsymbol{X}_t$, the visual content is encoded into vision tokens $\boldsymbol{H}_v$ by a vision encoder (*e.g.*, CLIP ViT-L/14 (Radford et al., 2021)) and a projection layer. Concurrently, the textual input $\boldsymbol{X}_t$ is converted into textual embeddings $\boldsymbol{H}_t$ via the LLM's embedding layer. These vision tokens and textual embeddings, concatenated into $\boldsymbol{H}$, are subsequently processed by the LLM (*e.g.*, Vicuna (Chiang et al., 2023)) to generate the output.

More precisely, the feed-forward network (FFN) sub-layer within the i-th layer of the LLM back-

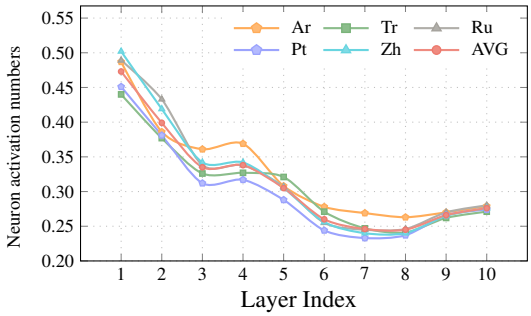


Figure 3: The number of activated neurons R_l^i across all non-English languages. "AVG" indicates the average activation level computed over these languages.

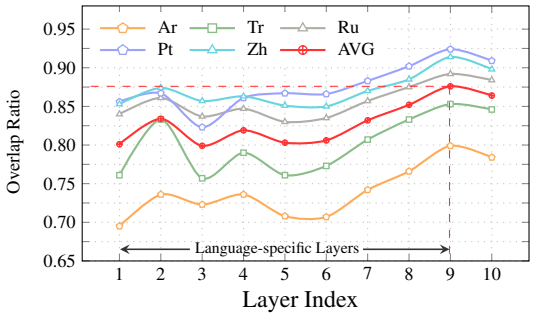


Figure 4: The overlap ratio O_l^i between non-English and English activated neurons. "AVG" indicates the average overlap ratio among all non-English languages.

bone is formulated as:

$$\text{FFN}(\boldsymbol{H}^i) = \left[f(\boldsymbol{W}^i_{\text{gate}}\boldsymbol{H}^i) \otimes (\boldsymbol{W}^i_{\text{up}}\boldsymbol{H}^i)\right] \boldsymbol{W}^i_{\text{down}}, \tag{1}$$

where $\boldsymbol{W}^i_{\text{gate}}, \boldsymbol{W}^i_{\text{up}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{inter}}}$ are weight matrices for the gate and up-projection respectively, and $\boldsymbol{W}^i_{\text{down}} \in \mathbb{R}^{d_{\text{inter}} \times d_{\text{model}}}$ is the down-projecting matrix. $f(\cdot)$ denotes a non-linear activation function.

In the context of an FFN sub-layer, a neuron is defined as a single column within $\boldsymbol{W}^i_{\text{up}}$, implying that each FFN sub-layer contains d_{inter} neurons. A specific neuron in the i-th FFN sub-layer is considered *activated* if its corresponding value within the $f(\boldsymbol{W}^i_{\text{gate}}\boldsymbol{H}^i)$ exceeds zero (Tang et al., 2024).

2.2 Experimental Setups for Pilot Study

To investigate the influence of language-specific neurons on the multilingual understanding capabilities of LVLMs, we conduct our pilot study using the MMBench dataset (Sun et al., 2025). MMBench is a comprehensive multilingual VQA dataset, featuring image-question pairs annotated across six distinct languages. For our analysis, we sample n instances per language. We denote the language of the question in each pair as l. For each language l, we quantify the activation of neurons by the image-question pair in the i-th layer as follows:

$$\boldsymbol{A}^i_l = \mathbb{I}[f(\boldsymbol{W}^i_{\text{gate}}\boldsymbol{H}^i) > 0], \tag{2}$$

where $\mathbb{I}$ is the element-wise indicator function. We then normalize the number of activated neurons by the total number of neurons in layer i to compute the proportion of activated neurons as follows:

$$R^i_l = \frac{\sum_{k=1}^{d_{\text{inter}}} (\boldsymbol{A}^i_l)_k}{d_{\text{inter}}}. \tag{3}$$

To analyze language-specific activation patterns, we categorize activated neurons sets $\mathcal{N}^i$ in layer i by English and non-English. Let $\mathcal{L}_{\text{non-eng}}$ denote the set of all non-English languages. The overlap ratio of activated neuron sets between each non-English $l \in \mathcal{L}_{\text{non-eng}}$ and English is computed as:

$$O^i_l = \frac{|\mathcal{N}^i_l \cap \mathcal{N}^i_{\text{eng}}|}{|\mathcal{N}^i_{\text{eng}}|}. \tag{4}$$

2.3 Observations

Our analysis reveals significant language-specific neurons activity throughout the decoder layers of LVLMs. Illustrative data for the LLaVA-1.5-7B are presented in Figure 3, which shows the count of activated neurons per layer, and Figure 4, which depicts the overlap ratio of activated neurons between non-English and English. We observe a progressive decline in the total number of activated neurons across non-English languages with increasing layer depth. Conversely, the overlap ratio of activated neurons between non-English and English increases, ultimately reaching a distinct peak.

This pattern suggests that language representations are initially distinct and independent, while progressively converging toward an English-centric pattern at greater depths. This confirms that **certain layers within the decoder are primarily responsible for processing language-specific representation**. Building upon this, we classify the layers preceding the point where the average overlap ratio among all non-English languages attains its maximum as *language-specific layers*. Full visualization of all layers is available in Appendix A.

3 Methodology

Building upon our findings, we propose a training method designed to achieve *efficient* multilingual enhancement of LVLMs, which includes two processes: (1) identifying layers responsible for multi-

lingual understanding; (2) precisely fine-tuning the selected layers to enhance multilingual capabilities.

3.1 Select Multilingual Understanding Layers

Prior research (Zhang et al., 2025a) suggests that LVLMs predominantly extract crucial information from visual representations in shallow to intermediate layers to facilitate the core task-solving process. Indiscriminately fine-tuning all language-specific layers might compromise inherent general capabilities embedded within them. Consequently, we introduce a more granular layer selection algorithm designed to strike a balance between enhancing multilingual understanding and maintaining the model's general abilities. To this end, we utilize the mean squared deviation (MSD) to precisely measure the stability of neuron activation across different languages. For each layer i within language-specific layers $\mathcal{K}$, we compute its MSD^i as follows:

$$\mu^i = \frac{1}{|L|} \sum_{l \in L} R_l^i, \tag{5}$$

$$\mathrm{MSD}^i = \frac{1}{|L|} \sum_{l \in L} (R_l^i - \mu^i)^2, \tag{6}$$

where L denotes the collection of all languages, and R_l^i represents the normalized count of activated neurons, as calculated by Equation (3).

A higher MSD^i for a given layer indicates a greater divergence in its activation pattern across different languages, suggesting a more substantial engagement in multilingual understanding rather than general abilities. To quantify the average engagement in multilingual understanding across language-specific layers, we calculate as follows:

$$\theta = \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} \mathrm{MSD}^i. \tag{7}$$

Layers whose MSD^i exceeds the threshold θ are ultimately selected for fine-tuning, as these layers contribute more significantly to multilingual understanding than to general capabilities.

3.2 Supervised Fine-Tuning for Enhancing Multilingual Understanding

Recent studies (Basu et al., 2024; Zhang et al., 2024a; Ye et al., 2025) have shown that text tokens effectively integrate visual information from vision tokens via attention sub-layers, particularly within the shallow layers of LVLMs. In light of this, we undertake fine-tuning across all modules within the selected layers. This not only preserves the model's fundamental ability to process visual information within these crucial layers, but also significantly enhances its capacity to comprehend multilingual questions. Given the visual inputs X_v, non-English questions $X_{t,l}$ and their English counterparts $X_{t,eng}$, the loss function is formulated as:

$$\mathcal{L} = - \sum_{l \in \mathcal{L}_{\text{non-eng}}} \log P(X_{t,\text{eng}} \mid X_{t,l}, X_v; \theta), \tag{8}$$

where θ represents the parameters of the selected layers actively involved in the fine-tuning process.

4 Experiment Settings

4.1 Datasets

To assess the efficacy of PLAST, our main experiments are conducted on two multilingual VQA datasets covering six languages, including Arabic (Ar), Turkish (Tr), Russian (Ru), Portuguese (Pt), Chinese (Zh), and English (En).

MMBench (Sun et al., 2025) consists of data in six languages, translated from the original MMBench (Liu et al., 2024c) using GPT-4 (OpenAI, 2023). These translations are subsequently verified manually to ensure their accuracy.

MMMB (Sun et al., 2025) is compiled by sampling items from the ScienceQA (Lu et al., 2022), MME (Fu et al., 2023), and SEED-Bench (Li et al., 2024) datasets, which are then translated into five other languages using GPT-4.

4.2 Evaluation Metrics

We follow the evaluation settings from Sun et al. (2025), primarily focusing on *accuracy*. During our assessments, we employ the VLMEvalKit from OpenCompass (Contributors, 2023), and ensure consistent configuration settings across all compared methods to facilitate a fair comparison.

4.3 Baselines

We compare our method with two types of baselines: *prompting-based* and *training-based*. To validate the generalizability of PLAST, we select three representative LVLMs with different sizes for evaluation: LLaVA-1.5-7B/13B (Liu et al., 2024a) and LLaVA-1.6-7B/13B (Liu et al., 2024b), and Qwen-VL-Chat (Bai et al., 2024). Comprehensive baseline details are provided in Appendix B.

Method	Training Cost	Training Layers	Trained Param.	MMBench							MMMB						
				Ar	Tr	Ru	Pt	Zh	En	Avg.	Ar	Tr	Ru	Pt	Zh	En	Avg.
LLaVA-1.5-7B				34.6	42.4	54.8	61.1	58.1	64.7	52.6	41.7	43.1	55.1	**59.2**	**57.7**	66.2	53.8
+ ITP	-	-	-	22.2	39.0	49.6	55.6	48.1	64.7	46.5	30.6	33.6	42.5	46.8	45.7	66.2	44.2
+ ETP	-	-	-	35.2	40.6	58.1	58.0	52.7	**64.7**	51.6	42.4	45.3	57.5	58.7	57.6	**66.2**	54.6
+ M-SFT	7.3×	1-32	100.0%	39.3	50.2	54.9	57.6	58.4	63.7	54.0	44.3	47.3	56.4	58.4	55.7	63.4	54.2
+ QALIGN	2.8×	1-32	100.0%	24.3	29.0	39.7	39.7	38.7	44.6	36.0	36.3	40.7	43.8	41.3	41.2	49.6	42.2
+ PLAST	1.0×	1-5	15.6%	**44.4**	**51.9**	**58.4**	**62.3**	**59.4**	64.2	**56.8**	**46.7**	**50.1**	**59.4**	57.1	56.8	65.1	**55.8**
LLaVA-1.5-13B				46.6	53.2	61.6	63.0	**63.2**	69.0	59.4	45.9	50.7	62.6	61.7	61.6	69.8	58.7
+ ITP	-	-	-	43.9	53.5	**62.4**	**64.7**	61.0	69.0	59.1	44.5	49.3	62.2	60.9	55.4	69.8	57.0
+ ETP	-	-	-	42.8	49.4	60.2	59.5	61.0	**69.0**	57.0	45.0	50.9	**62.9**	58.5	**63.8**	**69.8**	58.5
+ M-SFT	12.5×	1-40	100.0%	48.4	**59.5**	60.6	61.7	60.9	67.1	59.7	48.7	53.0	60.1	**63.1**	59.9	68.1	58.8
+ QALIGN	4.9×	1-40	100.0%	45.9	57.4	59.1	61.8	59.2	66.9	58.4	47.5	47.2	56.8	55.5	54.6	62.7	54.0
+ PLAST	1.0×	1-5	12.5%	**51.5**	58.7	62.2	64.3	62.3	67.6	**61.1**	**49.7**	**53.1**	61.8	61.5	60.6	69.2	**59.3**
Qwen-VL-Chat				36.7	40.1	47.9	49.1	56.0	57.3	47.9	43.0	44.1	51.7	46.4	57.8	56.0	49.8
+ ITP	-	-	-	23.5	37.2	42.6	44.8	49.5	57.3	42.5	33.2	34.6	39.3	34.4	46.0	56.0	40.6
+ ETP	-	-	-	38.3	41.7	49.5	48.2	53.6	57.3	48.1	45.9	47.9	52.8	45.5	53.6	56.0	50.3
+ M-SFT	7.3×	1-32	100.0%	41.6	**51.2**	48.7	52.4	58.1	58.2	51.7	47.4	48.2	54.6	48.5	**58.0**	58.5	52.5
+ QALIGN	2.8×	1-32	100.0%	22.3	31.6	31.5	34.7	38.2	37.4	32.6	32.0	39.1	45.3	36.9	41.4	41.2	39.3
+ PLAST	1.0×	1-5	15.6%	**47.9**	50.6	**51.8**	**54.6**	**58.7**	**59.0**	**53.8**	**48.4**	**52.6**	**54.8**	**50.2**	56.1	**60.7**	**53.8**

Table 1: The Accuracy (%) on the MMBench and MMMB benchmarks. "Avg." denotes the average accuracy across six languages. "Training Cost" refers to the time required to train the models. "Training Layers" specifies the decoder layers selected for trainig. "Trained Param." indicates the proportion of trainable parameters in the LLM backbone. **Bold** and underline numbers indicate the best performance and second performance among each group.

4.3.1 Prompting-based Methods

We select two established prompting strategies that enhance the multilingual capabilities of LVLMs via implicit and explicit translation, respectively.

Implicit Translation Prompting (ITP) (Shao et al., 2024). To enhance the model's ability across multilingual scenarios, we prompt the model to *implicitly translate* non-English questions into English, enabling it to *think in English*.

Explicit Translation Prompting (ETP) (Qin et al., 2023). This approach incorporates a two-stage prompting strategy: first prompting the model to *explicitly translate* non-English questions into English, then solving the task in English. This explicit translation mechanism has shown effectiveness in improving multilingual performance.

4.3.2 Training-based Methods

Multilingual Supervised Fine-tuning (M-SFT) further fine-tunes the visual instruction-tuned model with multilingual visual-instruction data, aiming to achieve better multilingual alignment.

QALIGN (Zhu et al., 2024) incorporates a question alignment stage, where the model is first trained to translate non-English image-question pairs into English. Then, the model is further fine-tuned using English-only visual-instruction data, effectively leveraging the acquired language translation capabilities for visual instruction alignment.

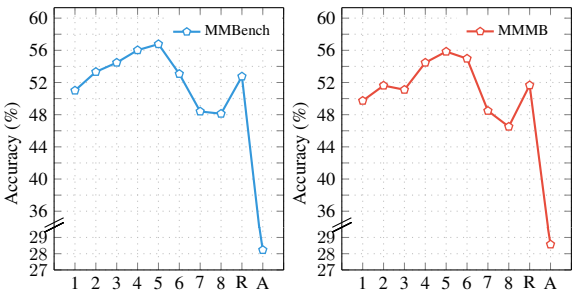
4.4 Implementation Details

Due to the scarcity of multilingual visual instruction training data, we first sample English image-text pairs from the ShareGPT4V dataset (Chen et al., 2024) and then translate the English questions into five other languages using GPT-4 to construct our training data. During the fine-tuning process, we only train the first five decoder layers of the visual instruction-tuned models. For more details, please refer to Appendix C.

5 Experimental Results

We present the main results of LLaVA-1.5-7B/13B and Qwen-VL-Chat on MMBench and MMMB benchmarks in Table 1. The results of LLaVA-1.6-7B/13B are presented in Table 7 in Appendix D.

LVLMs exhibit significant performance imbalances in multilingual scenarios. As shown in Table 1, while the LLaVA-1.5-7B model achieves a strong performance of 64.7 on the MMBench and 66.2 on the MMMB in English, its performance significantly declines in moderately low-resource languages, such as Arabic. Specifically, it shows a decrease of **46.5%** (64.7 → 34.6) on the MMBench and **37.0%** (66.2 → 41.7) on MMMB. These performance imbalances are notably exacerbated in prompting-based strategies. Both implicit and explicit prompting strategies struggle to improve performance and can even result in substantial decline. Notably, ITP results in an average performance de-



(a) End Training Layer (b) End Training Layer

Figure 5: Average accuracy across different training layers. The x-axis signifies training the decoder layers from the first up to the specified layer. "R" denotes randomly selected layers, and "A" denotes all layers.

crease of **21.7%** in Arabic and **10.2%** in Turkish among the three models, primarily due to the translation inaccuracies triggered by the model's inferior performance in relatively low-resource languages.

PLAST enhances multilingual performance and shows robust cross-model generalization. It can be observed that PLAST consistently outperforms all baselines on both MMBench and MMMB, achieving *state-of-the-art* performance. Specifically, it achieves an average improvement of 8.0% and 4.0% on MMBench and MMMB respectively across the three evaluation models. These results highlight the effectiveness of PLAST in enhancing the multilingual capabilities of LVLMs. Remarkably, PLAST leads to substantial gains for moderately low-resource languages, with improvements of **12.0%** (41.7 → 46.7) in Arabic and **16.2%** (43.1 → 50.1) in Turkish on the MMMB for the LLaVA-1.5-7B model. This suggests that PLAST, by leveraging only a modest amount of question-translation data, effectively bridges language gaps to some extent. In addition, among all evaluated models of varying architecture and scales, PLAST consistently delivers improvements, highlighting its strong generalizability across different models.

PLAST demonstrates superior efficiency. Unlike training-based baselines that require extensive multilingual image-text data for full-parameter fine-tuning, PLAST achieves remarkable efficiency by utilizing less than half of the training data (see Table 10) and selectively fine-tunes only shallow decoder layers. As detailed in Table 1, PLAST fine-tunes merely **15.6%** and **12.5%** of the LLM backbone's parameters in 7B and 13B models, respectively, resulting in a reduction of the training time by **7.3**× and **12.5**×. Additionally, compared to the two-stage QALIGN, PLAST effectively mitigates catastrophic forgetting during continual fine-tuning by precisely targeting the layers responsible for multilingual understanding. This targeted fine-tuning leads to significant improvements, with average performance gains of **42.5%** and **26.3%** on MMBench and MMMB benchmarks, respectively.

Method	Hi	Iw	Ro	Th	Avg.
LLaVA-1.5-7B	5.8	7.0	21.9	14.6	12.3
+ ITP	1.7	3.5	2.8	12.3	5.1
+ ETP	9.5	6.7	4.8	8.6	7.4
+ M-SFT	8.3	3.0	34.3	18.9	16.1
+ QALIGN	9.8	4.3	19.8	12.7	11.7
+ PLAST (Ours)	**10.7**	**10.2**	**36.7**	**20.8**	**19.6**

Table 2: The accuracy (%) on the MaXM benchmark.

6 Ablation Study and Further Analysis

We conduct extensive ablation studies and analysis to verify the effectiveness of PLAST. For more analysis, please refer to Appendix E.

Effect of layer selection strategy. To validate the necessity of the layer selection, we conduct ablation studies by training different layers within LVLMs. Our method identifies the top five layers as multilingual understanding layers by monitoring the dynamics of language-specific neurons using 100 parallel image-question pairs. As depicted in Figure 5, PLAST achieves the highest average accuracy on both the MMBench and MMMB. Training with an insufficient number of layers hinders the model's ability to understand multilingual questions effectively, whereas excessive layers impair the model's general capabilities, resulting in a decrease in accuracy. These results highlight that multilingual understanding is predominantly localized in the shallow layers of LVLMs. Precisely selecting layers that are actively involved in multilingual comprehension is essential for effectively enhancing the multilingual abilities of LVLMs without compromising their general abilities.

Extend to truly lower-resource languages. Given that MMBench and MMMB primarily encompass medium-resource languages, we further evaluate PLAST on MaXM benchmark (Changpinyo et al., 2023) to demonstrate its generalizability across a more diverse language families, particularly for low-resource languages. Specifically, we select Hindi (Hi), Hebrew (Iw), Romanian (Ro),

Method	Hi	Th	Ru	En	Avg.
LLaVA-1.5-7B	38.1	38.8	41.3	**47.8**	41.5
+ ITP	45.1	51.7	55.8	47.8	50.1
+ ETP	50.0	55.0	**59.2**	47.8	53.0
+ M-SFT	42.8	47.2	48.8	45.1	46.0
+ QALIGN	26.6	28.5	25.8	28.4	27.3
+ PLAST (Ours)	**56.9**	**59.5**	58.4	46.2	**55.3**

Table 3: The accuracy (%) on M5-VGR benchmark.

Method	Training Cost	Trained Param.	MMBench	MMMB
LLaVA-1.5-7B	-	-	52.6	53.8
+ LoRA (r=512)	4.9×	19.8%	52.3	53.2
+ PLAST	1.0×	15.6%	**56.8**	**55.8**
LLaVA-1.5-13B	-	-	59.4	58.7
+ LoRA (r=512)	6.5×	15.8%	57.9	55.4
+ PLAST	1.0×	12.5%	**61.1**	**59.3**

Table 4: The average accuracy of LoRA training strategy on the MMBench and MMMB test sets. For accuracy of all languages, refer to Table 8.

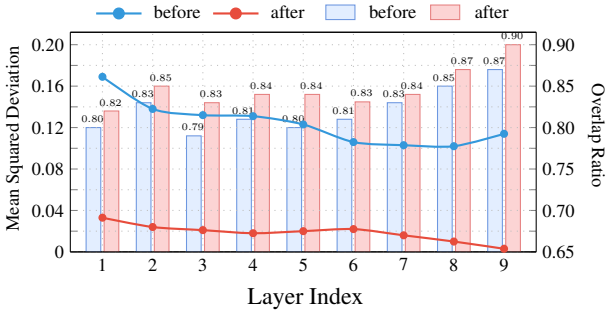


Figure 6: The comparison of the average overlap ratio (columns) and the MSD of activated neurons (curves) per layer in the LLaVA-1.5-7B model.

and Thai (Th) for evaluation, with results presented in Table 2. The empirical results reveal that PLAST achieves substantial performance improvements of 45.7% and 67.6% improvements for Iw and Ro, respectively, compared to the baseline models. These significant gains demonstrate the strong generalizability of our approach in truly low-resource settings. More details refer to Appendix F.1.

Generalization to complex reasoning tasks. To evaluate the cross-task transferability of PLAST, we conduct experiments on M5-VGR (Schneider and Sitaram, 2024), a challenging visually grounded reasoning benchmark spanning multiple languages. As illustrated in Table 3, PLAST shows substantial performance enhancements on complex reasoning tasks, particularly for low-resource languages, *e.g.*, Hi and Th, achieving an average improvement of 51.3% over the LLaVA-1.5-7B model. These findings confirm that the benefits of PLAST extend beyond fundamental VQA capabilities to more complex multimodal reasoning scenarios, highlighting its versatility and broader applicability across diverse multilingual task domains. For more experimental details, please refer to Appendix F.2.

PLAST indeed outperforms other parameter-efficient fine-tuning strategies. To further demonstrate the advantage of PLAST over other parameter-efficient approaches, we compare with the Low-Rank Adaptation (LoRA) (Hu et al., 2022; Zhong et al., 2025; Zhang et al., 2025b) strategy. Unlike PLAST that selectively fine-tunes all parameters within specific language-specific layers, LoRA trains a small subset of parameters across all layers. As shown in Table 4, with rank set to 512, PLAST requires *fewer* trainable parameters while simultaneously achieving *superior* performance, with improvements of 7.1% and 6.0% on MMBench and MMMB benchmarks, respectively. These results indicate the significance of language-specific layers, which precisely activate the model's multilingual capabilities, effectively preventing performance degradation associated with full-layer fine-tuning, while maintaining parameter efficiency. For more details, please refer to Appendix F.3.

Comparison of neuron activation and overlap ratio before and after training. To further investigate neuron activation dynamics before and after training, we compute the MSD^i using Equation (6) and the average overlap ratio across all non-English languages as specified in Equation (4). As shown in Figure 6, after training, the overlap ratio between non-English and English activated neurons shows a trend of an initial increase, followed by oscillating in the middle, and continuing to rise, which is consistent with the trend before training, representing that the model's representations gradually align with English representations. Notably, the overlap ratio across layers is significantly higher after training, indicating that PLAST effectively facilitates the transition from language-specific representations to English-centered representations in shallow layers. Moreover, the substantial decrease in MSD^i demonstrates that PLAST promotes more consistent neuron activation patterns, suggesting enhanced stability during multilingual processing.

Visual attention and semantic representation analysis before and after training. To assess the impact of multilingual alignment, we first uti-

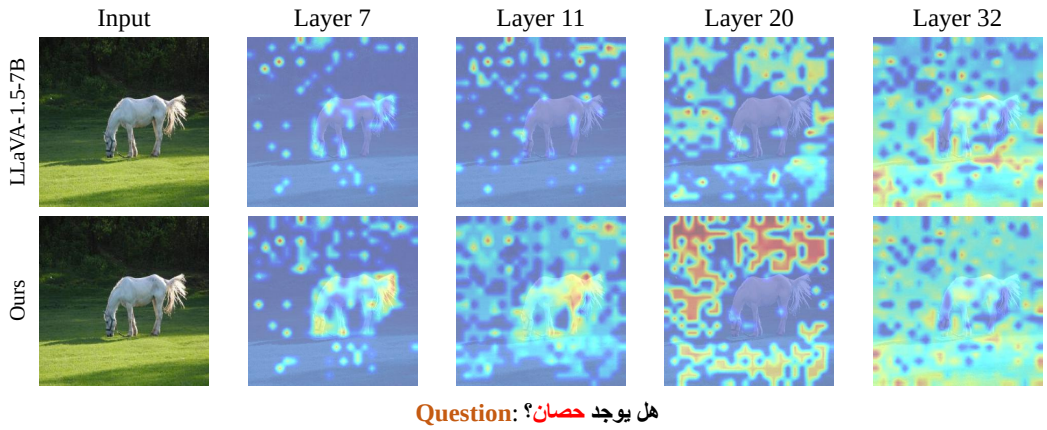


Figure 7: We compare the recognition of the object "horse" in images before and after training in LLaVA-1.5-7B using LLaVA-CAM (Zhang et al., 2025a), which reveals how attention scores guide the model to focus on relevant image regions during forward propagation based on the given questions. The case comes from the MMBench test sets, and the Arabic question in English means "**Is there any horse**?". For the visualization of the question in Turkish, and some qualitative case studies of PLAST and other baselines, please refer to the Appendix G.

lize LLaVA-CAM (Zhang et al., 2025a) to visualize the information flow from multilingual questions to corresponding image tokens. As shown in Figure 7, when processing an Arabic question explicitly referencing a specific visual element (*e.g.*, "horse"), models trained with PLAST demonstrate significantly enhanced attention allocation to the relevant regions in the 7^{th} and 11^{th} shallow layers. This indicates that by encouraging alignment between visual inputs and multilingual queries, PLAST enables the model to precisely capture language-specific visual features at early layers. Furthermore, we also investigate the transformation of linguistic representations through the model using t-SNE projections of language embeddings across various layers. Figure 20 (see Appendix F.4) reveals that the semantic space becomes notably more unified after training with PLAST, thereby enabling more effective capability sharing across languages.

7 Related Work

With the acceleration of globalization, multilingual LVLMs (Chen et al., 2023b; Li et al., 2023) have gained great attention for their ability to handle multiple languages comprehensively. However, due to the training corpora being mainly English-centric, these models perform significantly better in English than in other languages, leading to an imbalanced performance in multilingual scenarios.

Numerous approaches have been proposed to enhance the multilingual abilities of LVLMs, primarily categorized into prompting-based and training-based methods. Prompting-based methods leverage models' inherent understanding capabilities to translate non-English questions into English before generating final responses. For instance, Qin et al. (2023); Zhang et al. (2024b) employ either implicit or explicit prompting to guide the model to solve tasks step-by-step in English. Conversely, training-based methods focus on synthesizing multilingual training data via machine translation for multilingual visual instruction tuning (Sun et al., 2025; Maaz et al., 2024; Geigle et al., 2025). For example, Sun et al. (2025) constructs multilingual image-text pairs to train additional Mixture-of-Experts modules to convert English-biased features to language-specific features for multilingual alignment.

Unlike these approaches, our work focuses on achieving efficient multilingual capability enhancement. Rather than relying on extensive translated multilingual image-text pairs and full-parameter fine-tuning, PLAST precisely identifies the specific layers responsible for multilingual understanding, enabling more efficient multilingual alignment while maintaining superior performance.

8 Conclusion

This work proposes PLAST, a novel training recipe for efficient multilingual enhancement of LVLMs. PLAST first identifies layers predominantly engaged in multilingual understanding by monitoring language-specific neuron activations. These critical layers are then precisely fine-tuned with translation pairs to achieve multilingual alignment. Extensive evaluations show that PLAST significantly enhances LVLM multilingual performance.

Moreover, compared to full-parameter tuning methods, PLAST achieves superior performance and efficiency with only 14% of parameters tuned. Further analysis confirms its effectiveness across both low-resource languages and more complex visual reasoning tasks, demonstrating its broad applicability in diverse multilingual scenarios.

Limitations

This work exhibits several limitations worth noting. First, although our experiments cover a variety of model types (*e.g.*, the LLaVA series and Qwen-VL-Chat) and model scales (7B/13B), we do not conduct experiments on larger-scale models (larger than 13B) due to limited computing resources. We believe that PLAST still has great potential and is worth exploring on larger-scale models in future work. Second, while PLAST achieves comprehensive and efficient multilingual enhancement across on the MMBench, MMMB, MaXM, and M5-VGR benchmarks, the performance still involves trade-offs across different languages. We hypothesize that these trade-offs may arise from imbalances in multilingual training data during the pre-training and visual instruction tuning stages. However, our work starts from the perspective of language-specific layers to explore the efficient enhancement of multilingual abilities in LVLMs, rather than from the data-level. In future work, we will explore data-centric strategies for improving multilingual capabilities, especially data selection and data augmentation.

Ethics Statement

This work does not require ethical considerations. All the data used in this paper is sourced from open-source materials. Throughout the experimental process, all data and models were strictly utilized following their intended purposes and respective licenses. Additionally, this paper may contain offensive text related to the case study. We have all referenced them elliptically and will not present the complete harmful content within the paper.

Acknowledgements

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Fundamental Research Funds for the Central Universities, the Yunnan Fundamental Research Projects (No. 202401BC070021), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009). We would like to thank anonymous reviewers for their valuable comments.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. Understanding information storage and transfer in multi-modal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V. Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023a. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023b. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion

Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. Slam: Towards efficient multilingual reasoning via selective language alignment. In *International Conference on Computational Linguistics*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavavs. 2025. Centurio: On drivers of multilingual ability of large vision-language model. In *Proceedings of XYZ Conference*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. 2025a. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24564–24579. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, and Bing Qin. 2025b. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16896–16913. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024a. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14095–14113. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024b. Advancing large language model attribution through self-improving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3822–3836. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025c. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip (0.1).

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Muhammad Maaz, Hanoona Abdul Rasheed, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Tim Baldwin, Michael Felsberg, and Fahad Shahbaz Khan. 2024. PALO: A polyglot large multimodal model for 5b people. *CoRR*, abs/2402.14818.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10287–10299. Association for Computational Linguistics.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. LessWrong.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2695–2709. Association for Computational Linguistics.

Xiaoye Qu, Mingyang Song, Wei Wei, Jianfeng Dong, and Yu Cheng. 2024. Mitigating multilingual hallucination in large vision-language models. *CoRR*, abs/2408.00550.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *Preprint*, arXiv:2403.16999.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2025. Parrot: Multilingual visual instruction tuning. In *ICML*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *CoRR*, abs/2402.10588.

Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. 2025. CLAIM: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13080–13094, Vienna, Austria. Association for Computational Linguistics.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2025a. From redundancy to relevance: Enhancing explainability in multimodal large language models. *Annual*

Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics.

Xueyan Zhang, Jinman Zhao, Zhifei Yang, Yibo Zhong, Shuhao Guan, Linbo Cao, and Yining Wang. 2025b. UORA: Uniform orthogonal reinitialization adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11709–11728, Vienna, Austria. Association for Computational Linguistics.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2024a. Cross-modal information flow in multimodal large language models. *CoRR*, abs/2411.18620.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Jinman Zhao and Xueyan Zhang. 2024. Large language model is not a (multilingual) compositional relation reasoner. In *First Conference on Language Modeling*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yibo Zhong, Jinman Zhao, and Yao Zhou. 2025. Low-rank interconnected adaptation across layers. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 17005–17029. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8411–8423. Association for Computational Linguistics.

A Details of the number of activated neurons and overlap ratio for all models across all layers

We respectively present the complete visualization across all layers of LLaVA-1.5-7B/13B, LLaVA-1.6-7B/13B, and Qwen-VL-Chat in Figure 14-18.

B Experimental Details of baseline methods

We conduct experiments on the visual instruction-tuned LLaVA-1.5-Vicuna-7B/13B (Liu et al., 2024a) and LLaVA-1.6-Vicuna-7B/13B models (Liu et al., 2024b), and Qwen-VL-Chat. The LLaVA series models are equipped with the CLIP Vit-L/336px (Radford et al., 2021) as the vision encoder and the Vicuna-1.5 (Chiang et al., 2023) as the LLM backbone. And the Qwen-VL-Chat model is equipped with Openclip's ViT-bigG (Ilharco et al., 2021) as the vision encoder and Qwen-7B-Chat (Bai et al., 2023) as the LLM backbone. The detailed baseline implementation is as follows:

B.1 Prompting-based-methods

Implicit Translation Prompting (ITP) (Shao et al., 2024): For each question in the MMBench and MMMB test sets, ITP implicitly prompts LVLMs to first translate the non-English questions into English before reasoning in English. During evaluation, we utilize the VLMEvalKit from OpenCompass (Contributors, 2023) to evaluate the MMBench and MMMB test sets, and adopt the greedy decoding strategy for all models, setting the maximum generation length to 256. The evaluation prompt template used for MMBench and MMMB datasets is shown in Table 6. In the template, {*source_lang*} can be replaced with any of the following languages: Arabic, Turkish, Russian, Portuguese, Chinese, Hindi, Hebrew, Romanian, and Thai.

Explicit Translation Prompting (ETP) (Qin et al., 2023): For each question in the MMBench and MMMB test sets, ETP explicitly prompts LVLMs to first translate non-English questions into English, and then solve multimodal tasks with the translated questions. The evaluation prompt template used for MMBench and MMMB datasets is shown in Table 6. In the template, {*source_lang*} can be replaced with any of the following languages: Arabic, Turkish, Russian, Portuguese, Chinese, Hindi, Hebrew, Romanian, and Thai. {*Trans-*

Training Prompt	Translate this from [{*source_lang*}] to [English]:\n[{*source_lang*}]: {*source sentence*}\n[English]: {*English sentence*}

Table 5: The prompt used to train the decoder layers selected by PLAST.

MMBench & MMMB	**ITP Prompt**	Translate this question from [{*source_lang*}] to English and then answer with the option's letter from the given choices directly.
	ETP Prompt	Stage 1: Translate this question from [{*source_lang*}] to English. Stage 2: {*Translation result*}\nAnswer with the option's letter from the given choices directly.
MaXM	**ITP Prompt**	Translate this question from [{*source_lang*}] to English. \nAnd then only output the short answer in {language}:
	ETP Prompt	Stage 1: Translate this question from [{*source_lang*}] to English. Stage 2: {*Translation result*}\nAnd only output the short answer in {language}:
M5-VGR	**ITP Prompt**	Translate the question from [{*source_lang*}] to English. Based on the two images, is it correct? Yes or no? One word answer in English:
	ETP Prompt	Stage 1: Translate this question from [{*source_lang*}] to English. Stage 2: Based on the two images, is it correct to say {*Translation result*} ? Yes or no? One word answer in English:

Table 6: The prompt used for **ITP** and **ETP** baselines of MMBench, MMMB, MaXM, and M5-VGR test sets.

Method	**Training Cost**	**Training Layers**	**Trained Param.**	**MMBench**							**MMMB**						
				Ar	Tr	Ru	Pt	Zh	En	Avg.	Ar	Tr	Ru	Pt	Zh	En	Avg.
LLaVA-1.6-7B				37.2	46.0	57.9	62.3	**60.6**	68.0	55.3	40.5	44.5	62.4	60.6	**60.1**	70.1	56.4
+ ITP	-	-	-	10.6	29.3	41.6	44.0	49.4	68.0	40.5	27.2	31.0	42.4	41.6	31.7	70.1	40.7
+ ETP	-	-	-	35.3	40.2	**59.5**	58.7	56.1	**68.0**	53.0	43.5	47.1	**62.8**	**62.5**	59.9	**70.1**	57.7
+ M-SFT	7.3×	1-32	100.0%	41.7	**52.0**	56.6	61.8	59.6	65.6	56.2	47.4	47.9	59.8	59.4	58.3	68.4	56.8
+ QALIGN	2.8×	1-32	100.0%	35.7	43.3	53.1	54.1	51.1	58.8	49.4	40.3	42.1	51.2	47.9	47.7	59.6	48.1
+ PLAST	1.0×	1-5	15.6%	**43.3**	51.9	58.1	**63.0**	59.8	66.3	**57.1**	**50.5**	**48.5**	61.5	57.5	59.3	69.3	**57.8**
LLaVA-1.6-13B				45.4	52.9	**61.9**	64.1	**64.5**	70.9	59.9	45.4	50.6	**67.5**	**65.6**	66.8	73.5	61.6
+ ITP	-	-	-	23.7	36.2	49.5	43.8	58.2	70.9	47.0	31.3	28.7	40.3	36.3	46.6	73.5	42.8
+ ETP	-	-	-	42.0	49.3	60.5	62.9	60.3	**70.9**	57.7	49.0	**54.7**	67.3	62.7	**67.4**	**73.5**	**62.4**
+ M-SFT	12.5×	1-40	100.0%	47.3	54.1	60.8	64.7	63.4	69.7	60.0	50.9	51.4	67.1	64.2	64.9	72.4	61.8
+ QALIGN	4.9×	1-40	100.0%	39.2	49.9	53.6	55.2	51.7	59.5	51.5	48.8	48.3	57.2	51.5	52.2	62.9	53.5
+ PLAST	1.0×	1-5	12.5%	**49.4**	**56.3**	61.5	**65.1**	62.8	68.5	**60.6**	**53.7**	53.5	66.1	64.6	63.5	72.9	**62.4**

Table 7: The Accuracy (%) on the MMBench and MMMB benchmarks. "Avg." denotes the average accuracy across six languages. "Training Cost" refers to the time required to train the models. "Training Layers" specifies the decoder layers selected for training. "Trained Param." indicates the proportion of trainable parameters in the LLM backbone. **Bold** and underline numbers indicate the best performance and second performance among each group.

Method	**Training Cost**	**Training Layers**	**Trained Param.**	**MMBench**							**MMMB**						
				Ar	Tr	Ru	Pt	Zh	En	Avg.	Ar	Tr	Ru	Pt	Zh	En	Avg.
LLaVA-1.5-7B	-	-	-	34.6	42.4	54.8	61.1	58.1	**64.7**	52.6	41.7	43.1	55.1	**59.2**	**57.7**	**66.2**	53.8
+ LoRA (r=512)	4.9×	1-32	19.8%	37.5	48.3	53.1	56.2	56.9	62.0	52.3	43.8	46.6	55.2	56.3	54.6	62.8	53.2
+ PLAST	1.0×	1-5	15.6%	**44.4**	**51.9**	**58.4**	**62.3**	**59.4**	64.2	**56.8**	**46.7**	**50.1**	**59.4**	57.1	56.8	65.1	**55.8**
LLaVA-1.5-13B	-	-	-	46.6	53.2	61.6	63.0	**63.2**	**69.0**	59.4	45.9	50.7	**62.6**	**61.7**	**61.6**	**69.8**	58.7
+ LoRA (r=512)	6.5×	1-40	15.8%	45.8	54.3	58.5	61.9	61.2	65.9	57.9	45.8	49.3	54.2	59.8	58.1	65.4	55.4
+ PLAST	1.0×	1-5	12.5%	**51.5**	**58.7**	**62.2**	**64.3**	62.3	67.6	**61.1**	**49.7**	**53.1**	61.8	61.5	60.6	69.2	**59.3**
LLaVA-1.6-7B	-	-	-	37.2	46.0	57.9	62.3	**60.6**	**68.0**	55.3	40.5	44.5	**62.4**	**60.6**	**60.1**	**70.1**	56.4
+ LoRA (r=512)	4.9×	1-32	19.8%	41.6	48.9	56.2	61.0	59.4	66.4	55.9	46.3	44.8	59.1	57.0	57.9	66.0	55.2
+ PLAST	1.0×	1-5	15.6%	**43.3**	**51.9**	**58.1**	**63.0**	59.8	66.3	**57.1**	**50.3**	**48.5**	60.4	57.4	59.3	69.3	**57.5**
LLaVA-1.6-13B	-	-	-	45.4	52.9	**61.8**	64.1	**64.5**	**70.9**	59.9	45.4	50.6	**67.5**	**65.6**	**66.8**	**73.5**	61.6
+ LoRA (r=512)	6.5×	1-40	15.8%	45.7	52.4	56.8	63.8	63.1	68.1	58.3	49.3	48.4	66.2	61.9	60.4	72.1	59.7
+ PLAST	1.0×	1-5	12.5%	**49.4**	**56.3**	61.5	**65.1**	62.8	68.5	**60.6**	**53.1**	**53.5**	66.1	64.6	63.5	72.9	**62.3**

Table 8: The Accuracy (%) on the MMBench and MMMB test sets of LoRA training strategy.

Method	MMBench							MMMB						
	Ar	Tr	Ru	Pt	Zh	En	Avg.	Ar	Tr	Ru	Pt	Zh	En	Avg.
LLaVA-1.5-7B	34.6	42.4	54.8	61.1	58.1	**64.7**	52.6	41.6	43.1	55.1	**59.2**	**57.7**	**66.2**	53.8
+ PLAST	**44.4**	**51.9**	58.4	**62.3**	**59.4**	64.2	**56.8**	**46.7**	**50.1**	**59.4**	57.1	56.8	65.1	**55.8**
+ w/o MLP	39.2	45.4	56.1	58.1	55.8	61.9	52.8	44.5	46.8	57.2	50.4	54.9	63.6	52.9
+ w/o Attention	43.7	49.4	**58.6**	60.9	58.5	63.2	55.7	46.3	45.2	50.7	56.2	56.9	64.1	53.3

Table 9: The Accuracy (%) of training different sub-layers on MMMB and MMBench test sets across all languages.

lation result} can be replaced with the translation result by the model itself in the first round of dialogue.

B.2 Training-based-methods

Multilingual Supervised Fine-tuning (M-SFT): This method involves directly full-parameter fine-tuning models with multilingual instruction-following data during the visual instruction tuning stage. During training, we adopt the training hyperparameter from Liu et al. (2024a), using the M-ShareGPT4V (Sun et al., 2025) dataset. We keep the vision encoder frozen and train the projector and the LLM backbone for one epoch using eight NVIDIA A800 GPUs. The total batch size is set to 128, and the learning rate is maintained at 2e-5. The maximum input sequence length is set to 2,048 tokens. During decoding, we adopt the same decoding hyperparameter as the prompting-based method.

QALIGN: Zhu et al. (2024) proposes a two-stage training strategy to enhance multilingual abilities. In the first stage, we train the visual instruction-tuned model using multilingual question translation data paired with images, referred to as M-ShareGPT4V-Q, to translate non-English questions into corresponding English questions that convey the same meaning. In the second stage, to recover the general capabilities that were compromised during the first stage, we employ the English image-text instruction-following data paired with images, used in the first stage, referred to as ShareGPT4V-Sub, for visual instruction fine-tuning. In both two stages, we freeze the visual encoder layers and fine-tune the projection as well as all the decoder layers in the LLM backbone for one epoch using eight NVIDIA A800 GPUs. The total batch size is set to 128, and the learning rate is set to 2e-5. The maximum input sequence length is set to 2,048 tokens.

Function	Dataset	Usage	Lang	Size
Training	M-ShareGPT4V	M-SFT	Ar, Tr, Ru, Pt, Zh, En	138,000
	M-ShareGPT4V-Q	PLAST, QALIGN	Ar, Tr, Ru, Pt, Zh	60,000
	ShareGPT4V-Sub	QALIGN	En	12,000
Evaluation	MMBench	-	Ar, Tr, Ru, Pt, Zh, En	25,205
	MMMB	-	Ar, Tr, Ru, Pt, Zh, En	11,879
	MaXM	-	Hi, Iw, Ro, Th	1092
	M5-VGR	-	Hi, Th, Ru, En	478

Table 10: Dataset Statistics used for training and evaluation. "Usage" indicates the training data used by each method, "Lang" denotes the languages covered, and "Size" denotes the total number of samples. "M-ShareGPT4V" refers to the multilingual ShareGPT4V dataset (Sun et al., 2025), and "ShareGPT4V-Sub" is a subset sampled from ShareGPT4V (Chen et al., 2024). "M-ShareGPT4V-Q" contains question-translation data from "ShareGPT4V-Sub" and "M-ShareGPT4V".

C Experimental Details of PLAST

C.1 Training and Evaluation Datasets

To avoid translation errors caused by overly complex image-text pairs, we select the relatively simple sentence structures from the coco and GQA datasets within the ShareGPT4V dataset. Then we sort these data by length, extracting 7,200 and 4,800 image-text pairs from coco and GQA, respectively, which we refer to as ShareGPT4V-Sub. Subsequently, we translate the questions in ShareGPT4V-Sub into Arabic, Turkish, Russian, Portuguese, and Chinese using GPT-4 (OpenAI, 2023), followed by manual calibration. Finally, we use these translated questions paired with images to construct X-English question-translation data, referred to as M-ShareGPT4V-Q for PLAST training. We provide detailed statistics of the training set in Table 10, including the training data adopted by each method, the total number of samples, and the languages involved. For the main experiments, we select MMBench and MMMB as the evaluation benchmarks. Table 10 presents the data volumes and the languages included in MMBench and MMMB test sets. For the analysis experiment, we chose MaXM and M5-VGR as evaluation benchmarks. The specific data volume and included lan-

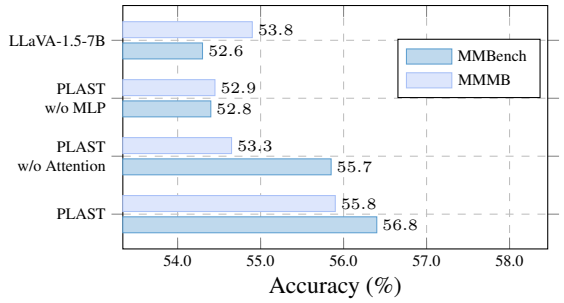


Figure 8: Comparison of the average accuracy of training different sub-layers on MMBench and MMMB. The detailed accuracy of all languages is shown in Table 9.

guages in these test sets are shown in Table 10.

C.2 Training Prompts

The prompt template employed for training is shown in Table 5. The prompt explicitly trains models to translate multilingual questions into English. In the template, {*source_lang*} can be replaced with any of the following languages: Arabic, Turkish, Russian, Portuguese, and Chinese. The placeholder {*source sentence*} is substituted with the multilingual questions, and {*English sentence*} is replaced with the corresponding English questions that convey the same meaning.

C.3 Training Details

We use LLaVA project[2] as our training framework. Training is conducted on eight NVIDIA A800 GPUs using Deepspeed stage 2 (Rajbhandari et al., 2020) for efficient multi-GPU distribution, with training precision set to Bfloat16. We maintain a total batch size of 128, a learning rate of 2e-5, and a maximum input sequence length of 2,048 tokens. Both the LLaVA-1.5-7B/13B and LLaVA-1.6-7B/13B models are trained over 2 epochs.

D Experimental Results of LLaVA-1.6

We present the complete results of LLaVA-1.6-7B and LLaVA-1.6-13B in Table 7.

E Additional Analysis

FFN sub-layers in LLMs have been recognized as storing the multilingual knowledge (Dai et al., 2022; Zhao et al., 2024; Tang et al., 2024). Therefore, we investigate the function of FFN and Attention sub-layers of selected layers by separately trainig them. As indicated in Figure 8, compared with only training the FFN and Attention sub-layers, PLAST further improves average accuracy by 5.1% and 4.8% across MMBench and MMMB test sets, respectively. This indicates that PLAST, by training the entire selected layer, not only improves the fusion of language information and visual information but also enhances the multilingual understanding abilities of LVLMs.

F Experimental Details of Ablation Studies and Analysis

F.1 Experimental Details of MaXM Benchmark

We follow the main experimental settings and utilize GPT-4 to translate the "ShareGPT4V-Sub" dataset into four low-resource languages included in the MaXM test set (Hindi, Hebrew, Romanian, and Thai) for mixed-language training. The evaluation prompt used for MaXM test set is shown in Table 6. To ensure a fair comparison and mitigate the risk of hallucinations (Huang et al., 2025c, 2024a, 2025b,a, 2024b) that can arise from complex multilingual questions, we use evaluation scripts[3] provided by Schneider and Sitaram (2024).

F.2 Experimental Details of Complex Reasoning Tasks

Following the main experimental settings, we utilized GPT-4 to translate the "ShareGPT4V-Sub" dataset into Hindi and Thai as included in the M5-VGR benchmark, and combined these translations with Russian training data from "M-ShareGPT4V" for mixed-language training. The evaluation prompt used for the M5-VGR test set is shown in Table 6. For a fair comparison, we use evaluation scripts[3] provided by Schneider and Sitaram (2024).

F.3 Experimental Details of LoRA Training Strategy

We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) as an alternative to **M-SFT**. For the LoRA training, we use a rank of 512, and the LoRA target modules are 'q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj'. We set a total batch size of 128, a learning rate of 2e-4 for the LLM backbone, a learning rate of 2e-5 for the projector, with a 0.03 linear warmup ratio, and a maximum input sequence length of 2,048 tokens. All the models are trained over 1 epoch.

[2] https://github.com/haotian-liu/LLaVA

[3] https://github.com/floschne/m5b

F.4 Experimental Details of t-SNE

To better understand the model's multilingual capability, we visualize the hidden state representations of the final token in multilingual questions, as it plays a crucial role in guiding the model's subsequent output (Wendler et al., 2024). Specifically, we randomly sample 250 instances from each of the five languages in the MMBench test set and extract 4096-dimensional hidden state representations from each layer of LLaVA-1.5-7B, both before and after training. These representations are then projected into a 2D space using t-SNE for visualization, as illustrated in Figure 20.

G Case Study

The visualization of attention scores for the question in Turkish is shown in Figure 19. Furthermore, we provide several qualitative examples from MMBench test sets in Figure 21 and Figure 22 to compare different methods of enhancing the multilingual abilities on LLaVA-1.5-7B. As shown in Figure 21, 22, the ITP often leads the model to produce mere translations under low-resource language conditions, failing to follow instructions to provide the final answer. In contrast, PLAST achieves efficient alignment of multilingual capabilities in LVLMs by facilitating shallow-layer understanding of multilingual instructions and integrating this information with visual features. In addition, we provide examples from the MaXM test sets in Figure 23 and 24 respectively. As shown in Figure 23 and 24, PLAST not only performs well in multilingual VQA types of multiple-choice questions, but also can follow instructions and answer correctly in multilingual non-multiple-choice question types.

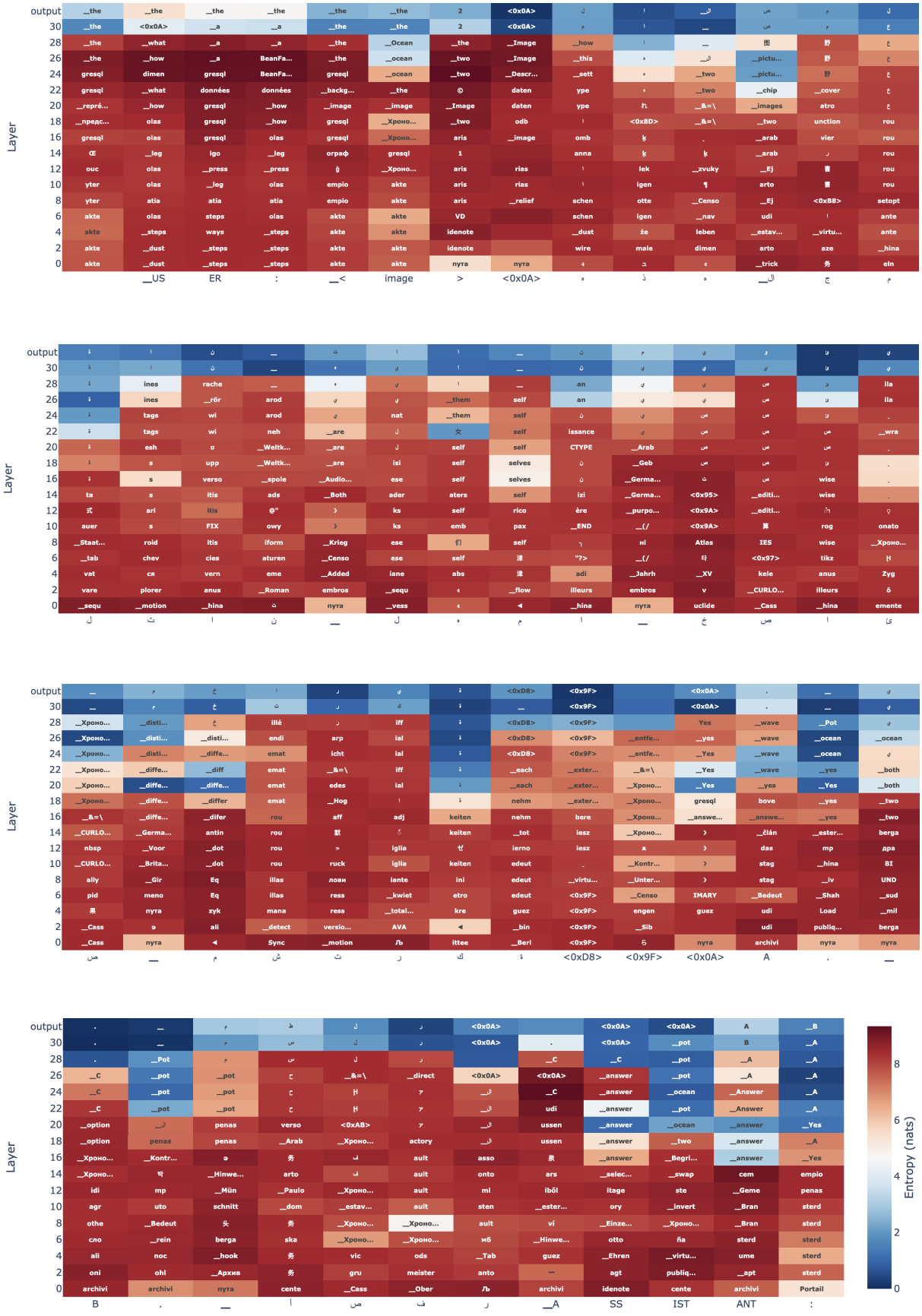


Figure 9: The complete visualization of next-token distributions in Arabic.

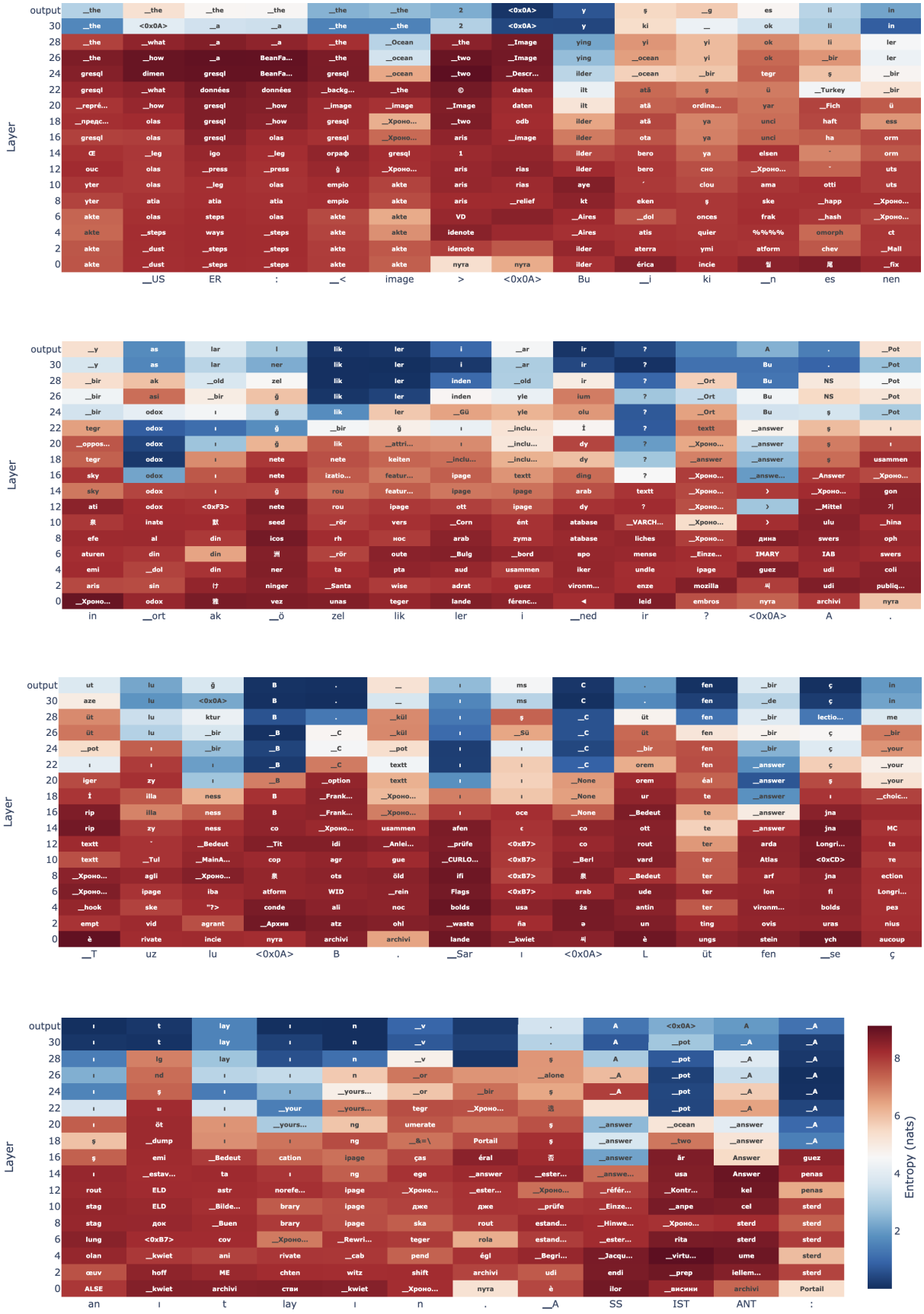


Figure 10: The complete visualization of next-token distributions in Turkish.

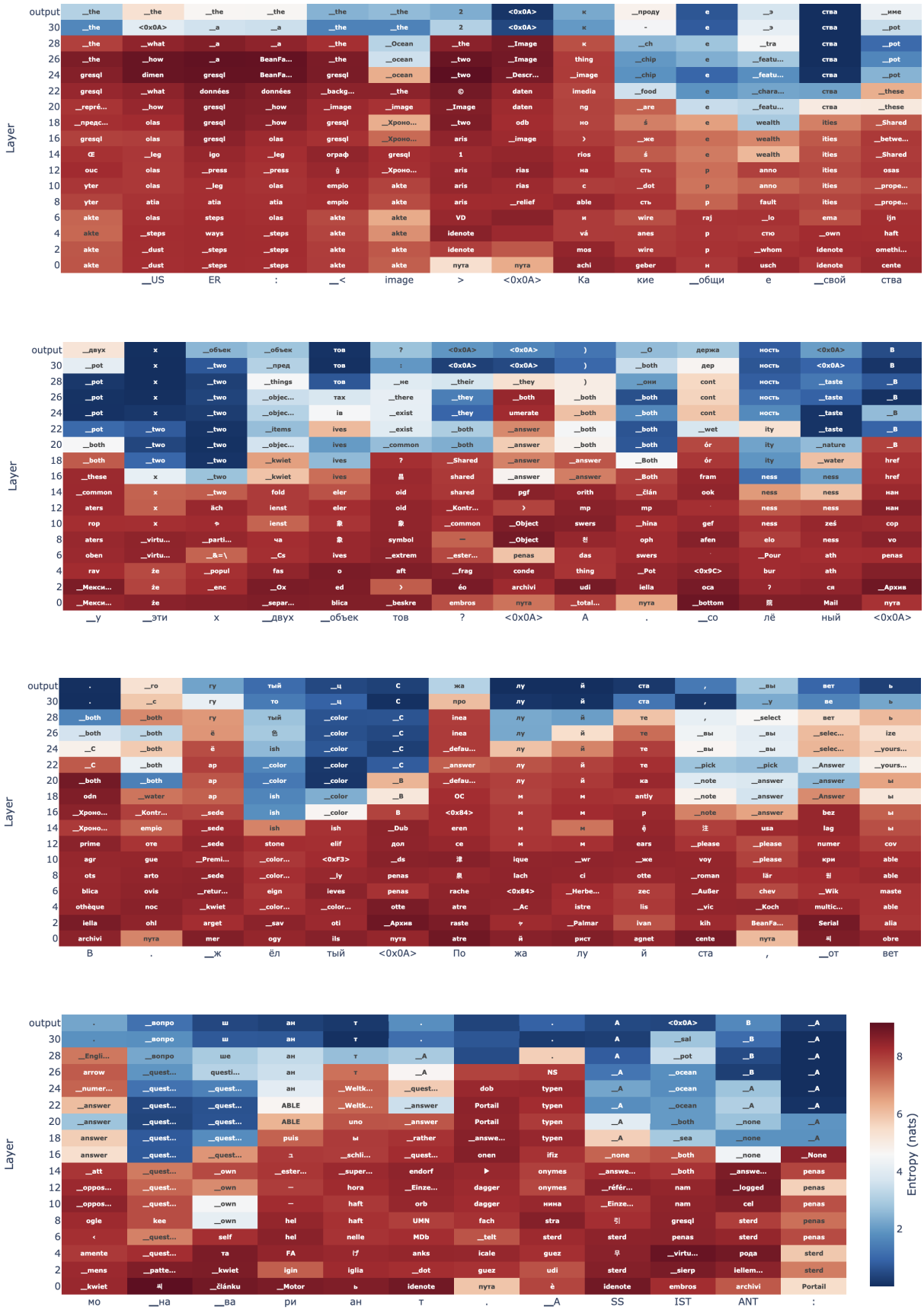


Figure 11: The complete visualization of next-token distributions in Russian.

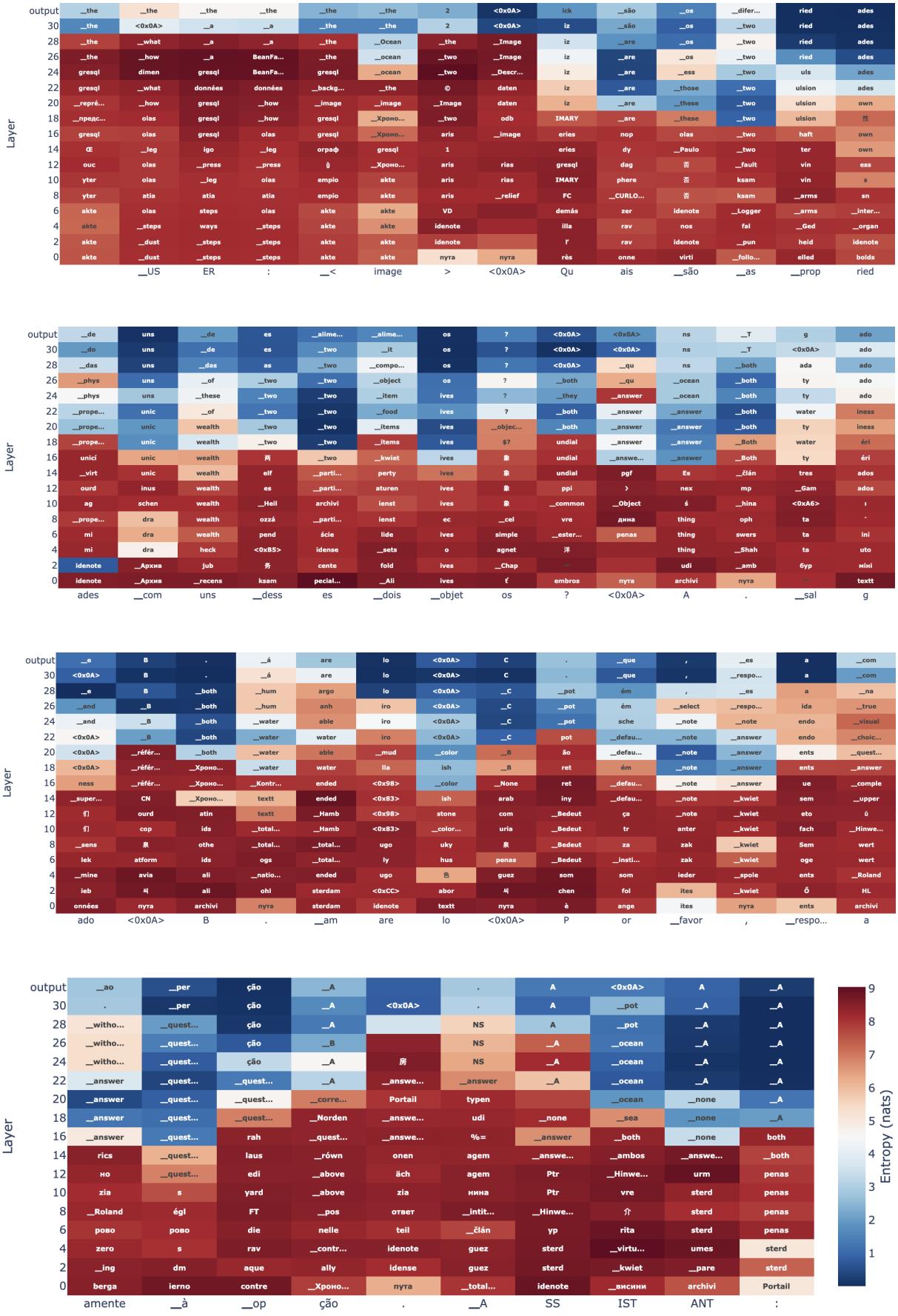


Figure 12: The complete visualization of next-token distributions in Portuguese.

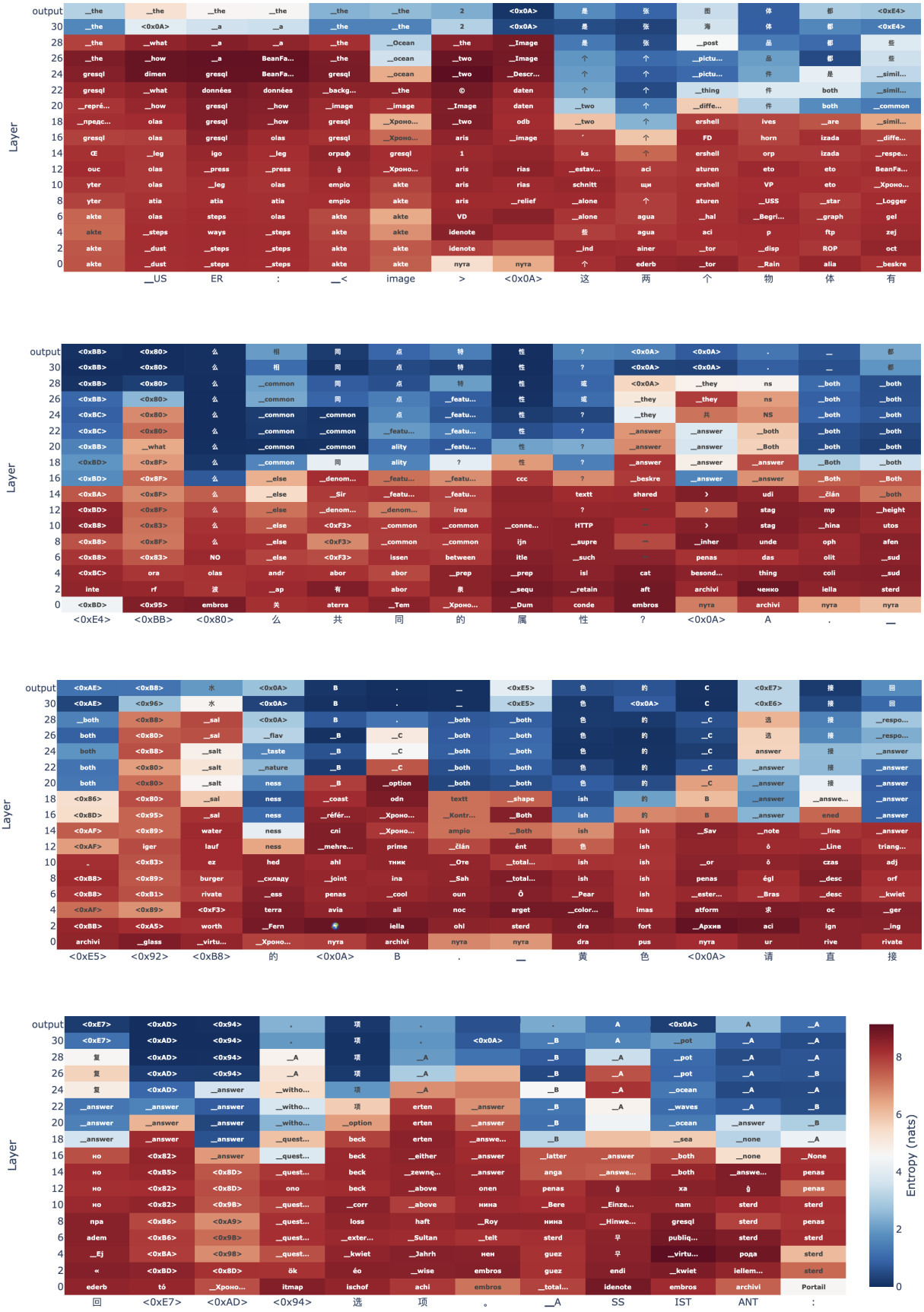


Figure 13: The complete visualization of next-token distributions in Chinese.

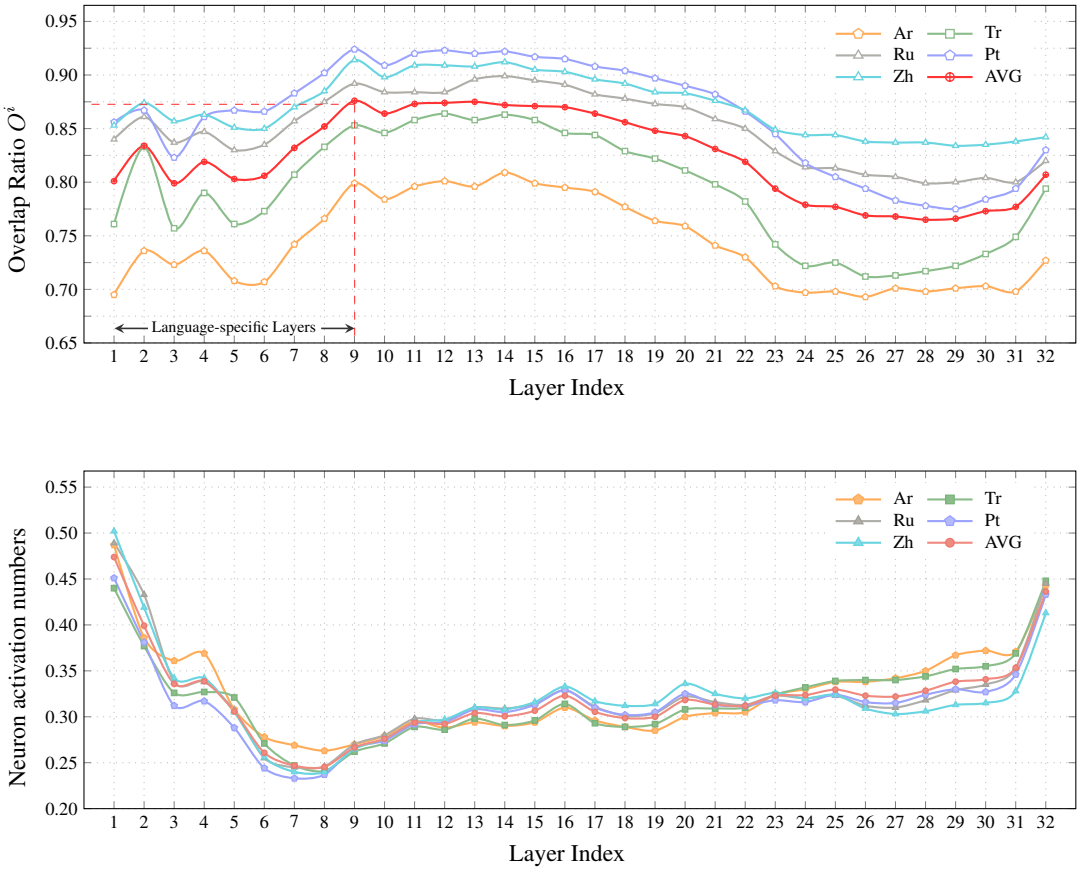


Figure 14: The overlap ratio between non-English and English activated neurons and the normalized number of activated neurons across non-English languages in the LLaVA-1.5-7B model.

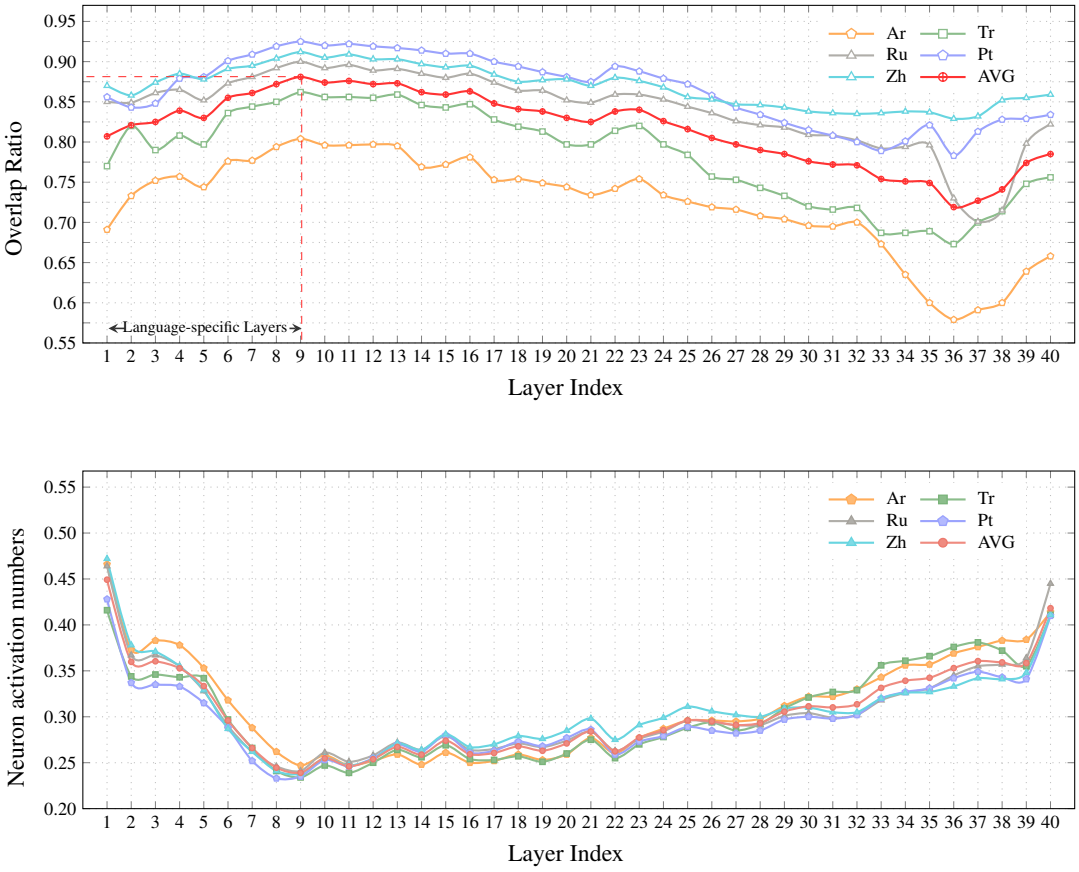


Figure 15: The overlap ratio between non-English and English activated neurons and the normalized number of activated neurons across non-English languages in the LLaVA-1.5-13B model.

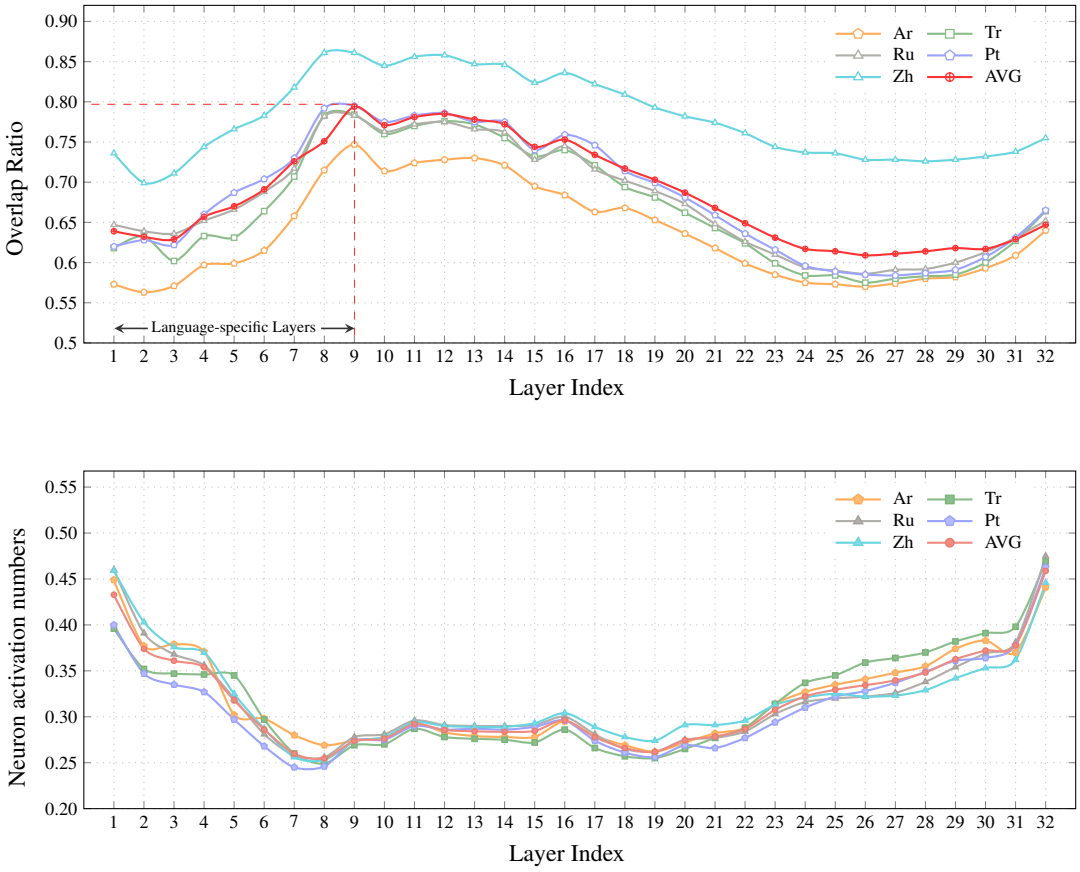


Figure 16: The overlap ratio between non-English and English activated neurons and the normalized number of activated neurons across non-English languages in the LLaVA-1.6-7B model.

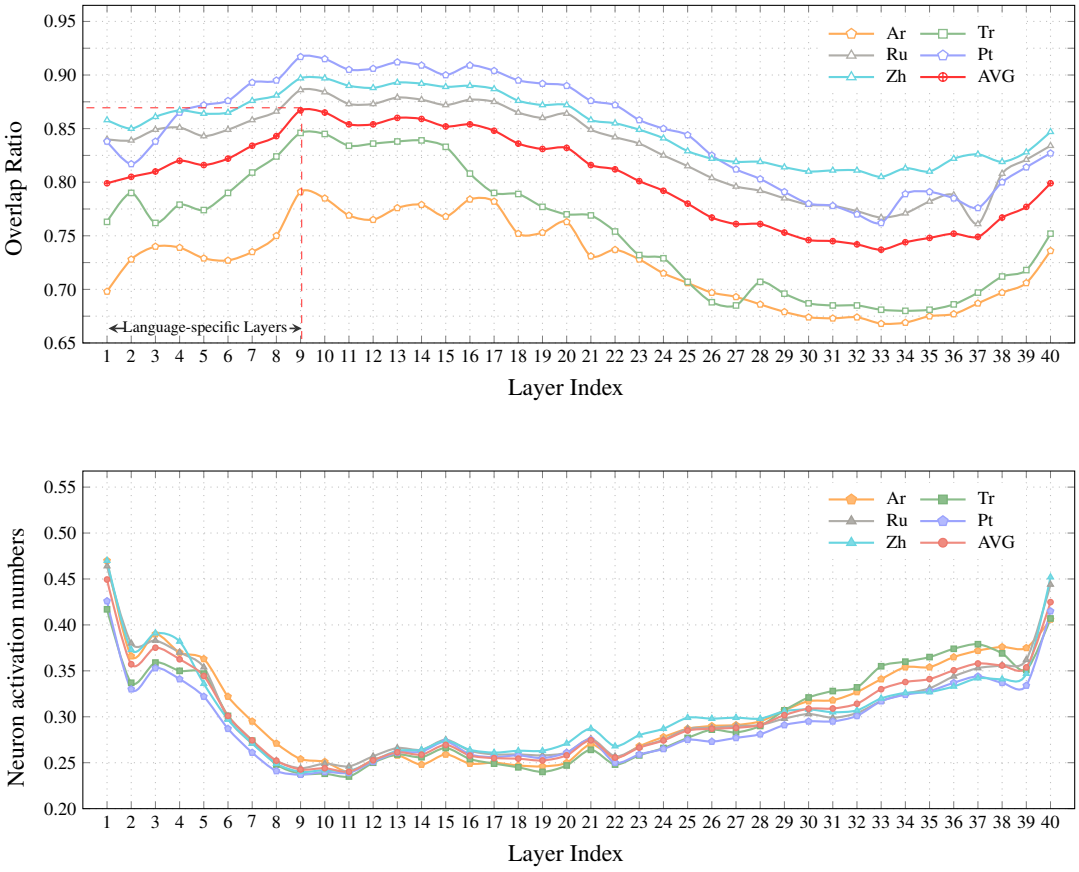


Figure 17: The overlap ratio between non-English and English activated neurons and the normalized number of activated neurons across non-English languages in the LLaVA-1.6-13B model.

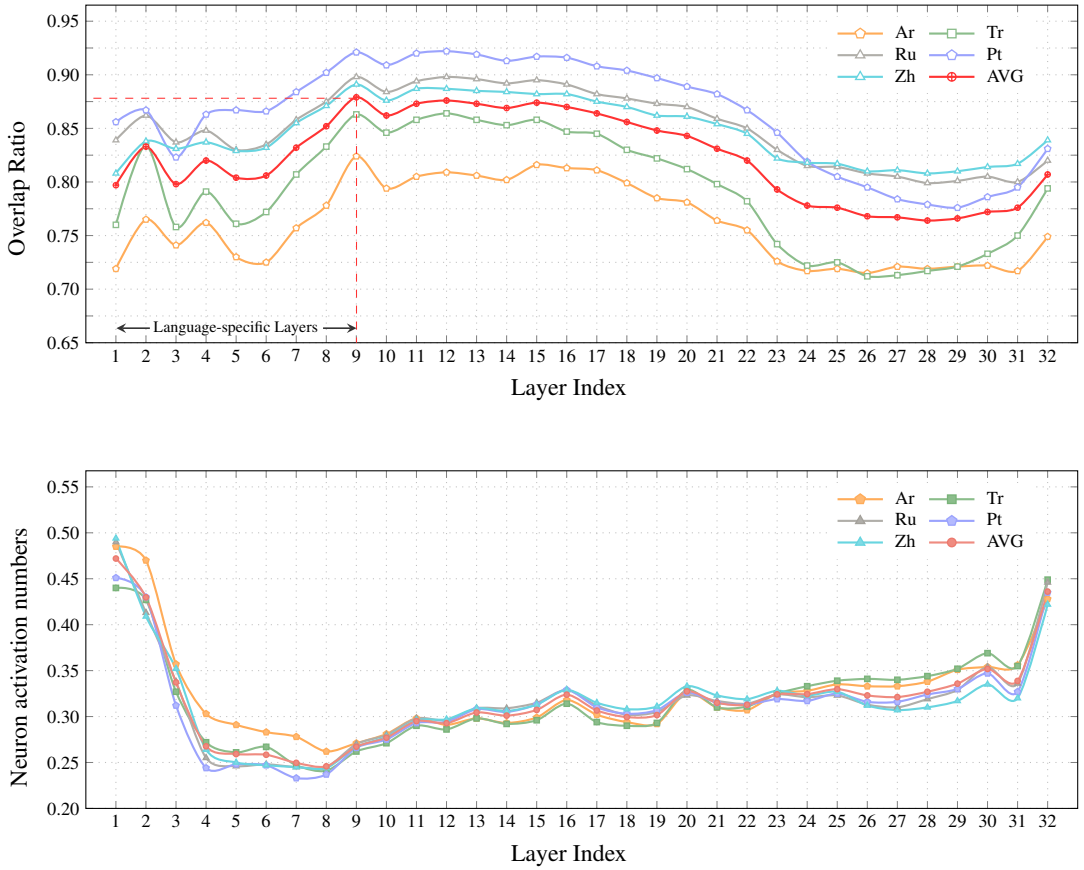


Figure 18: The overlap ratio between non-English and English activated neurons and the normalized number of activated neurons across non-English languages in the Qwen-VL-Chat model.

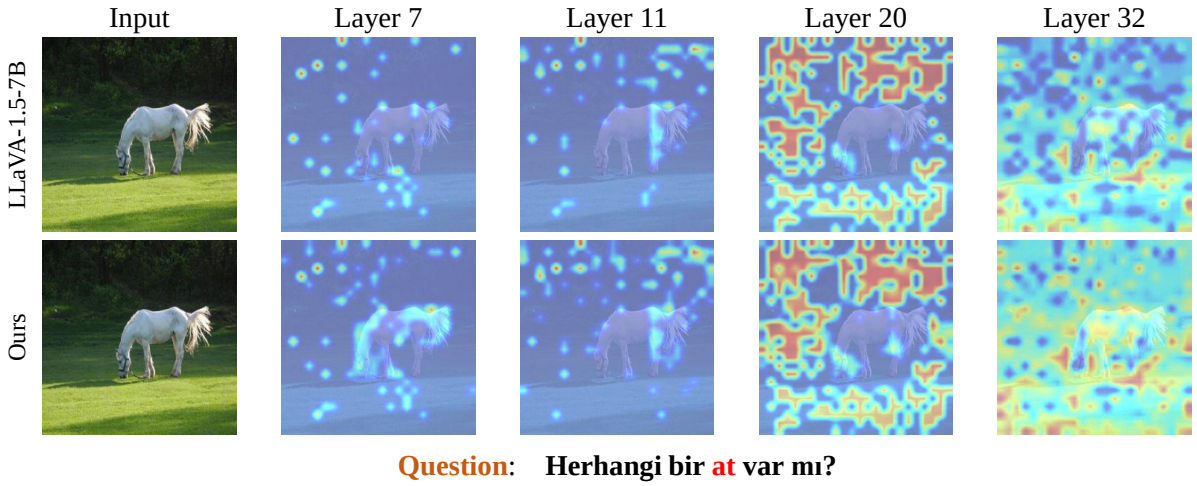


Figure 19: We compare the recognition of the object "horse" in images before and after training in LLaVA-1.5-7B using LLaVA-CAM (Zhang et al., 2025a), which reveals how attention scores guide the model to focus on relevant image regions during forward propagation based on the given questions. The case comes from the MMBench test sets, and the question in English means "**Is there any horse**?"

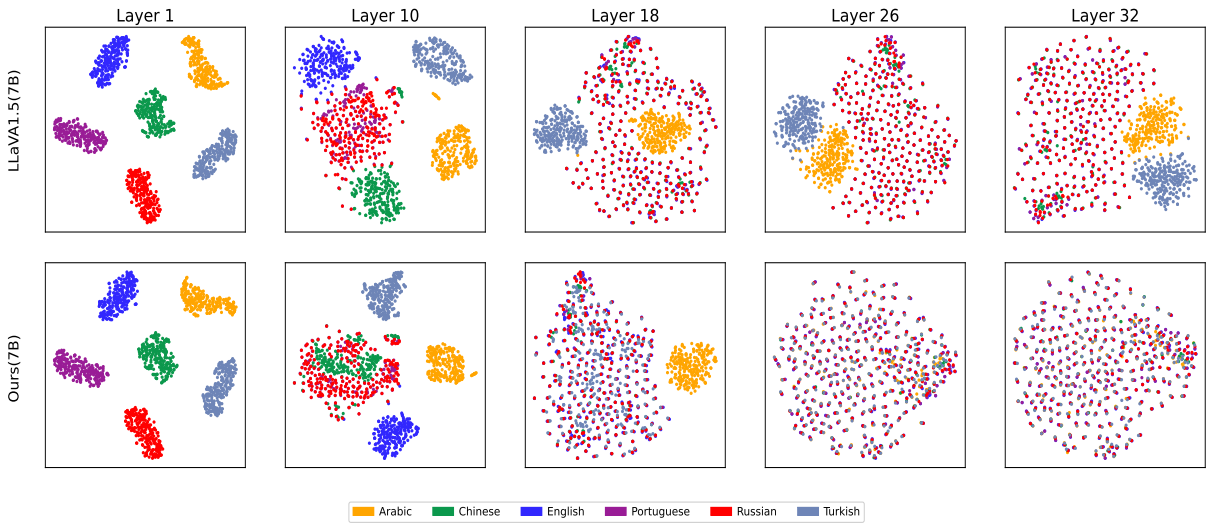


Figure 20: The final token representations from the MMBench test set are visualized using t-SNE for dimensionality reduction. The distributions in LLaVA1.5-7B at the 1^{th}, 10^{th}, 18^{th}, 26^{th}, and 32^{th} layers are compared before and after training. Different colors represent different languages.

Instruction:

<image>\n ما هو الكائن في هذه الصورة؟A. كوب B. حاوية القمامة C. صَحن D. طبق

LLaVA-1.5-7B:

B ✗

+ ETP:

What is the object in this picture? A. Cup B. Rubbish bin C. Plate D. Plate

B ✗

+ M-SFT:

B ✗

+ Ours:

A ✓

Figure 21: Example where PLAST can yield a correct answer compared to other baselines. The case comes from the MMBench Arabic test sets, and the question in English means "What is the object in this picture? A. Cup B. Trash can C. Bowl D. Plate".

Instruction:

<image>\n 土库曼斯坦在伊朗的哪个方向？ A. 东 B. 南 C. 西 D. 北

LLaVA-1.5-7B:

C ✗

+ ETP:

In which direction is Turkmenistan from Iran? A. East B. South C. West D. North

A ✗

+ M-SFT:

C ✗

+ Ours:

D ✓

Figure 22: Example where PLAST can yield a correct answer compared to other baselines. The case comes from the MMBench Chinese test sets, and the question in English means "In which direction is Turkmenistan from Iran? A. East B. South C. West D. North".

Instruction:
<image>\n कांच के गिलास में तरल किस रंग का है?

LLaVA-1.5-7B:
रंग का निर्देश नहीं है, क्योंकि ✗

+ ETP:
What is the cup made of?
Glass ✗

+ M-SFT:
टेबल ✗

+ Ours:
गुलाबी ✓

Figure 23: Example where PLAST can yield a correct answer compared to other baselines. The case comes from the MaXM Hindi test sets. The question in English means "What is the color of the liquid in the glass?", and the correct answer in English means "pink".

Instruction:
<image>\n מה יש על הגז שנדלק?

LLaVA-1.5-7B:
stove ✗

+ ETP:
What is on the gas stove?
Kettle ✗

+ M-SFT:
Fire ✗

+ Ours:
סיר עם מכסה ✓

Figure 24: Example where PLAST can yield a correct answer compared to other baselines. The case comes from the MaXM Hebrew test sets. The question in English means "What is on the lit stove?", and the correct answer in English means "A pot with a lid".