NeuPAT: Neuron-aware Plasticity Allocation Tuning for Language-Preserving MLLMs

Jiayue Jin[1,2], Jingwei Zhang[2,5], Chen Wang[2,6], Jing Liu[2,3,4], Longteng Guo[2,3,4*]

[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]Zhongguancun Academy, [3]Institute of Automation, Chinese Academy of Sciences
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences
[5]Tianjin University, [6]Nankai University

Correspondence: longteng.guo@nlpr.ia.ac.cn

Abstract

Multimodal expansion of large language models (LLMs) enables new perceptual capabilities but often compromises the language intelligence acquired during pretraining. In this work, we investigate this phenomenon from the perspective of internal adaptation dynamics and discover that neurons in pretrained LLMs exhibit heterogeneous plasticity during multimodal learning: some neurons are critical for preserving language capabilities, while others are more adaptive to multimodal knowledge. Based on this insight, we propose NeuPAT (Neuron-aware Plasticity Allocation Tuning), a lightweight and architecture-agnostic framework that allocates neuron-wise update constraints during multimodal instruction tuning. NeuPAT uses a small-scale probing stage to estimate neuron adaptation patterns and selectively protects language-sensitive neurons while promoting multimodal adaptation through more plastic neurons. Experiments across diverse LLM families demonstrate that NeuPAT recovers 94.5% of the language capability degradation caused by vanilla tuning on 11 language benchmarks while maintaining comparable multimodal performance, providing an effective approach for capability-preserving multimodal expansion.

1 Introduction

Large language models (LLMs) demonstrate strong capabilities in language understanding, reasoning, knowledge acquisition, and code generation through large-scale text pretraining (Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023). Extending these models beyond text has driven rapid progress in multimodal large language models (MLLMs) (Liu et al., 2023b; Dai et al., 2023; Wang et al., 2024; Bai et al., 2025; Li et al., 2024; Team et al., 2025; Wang et al., 2025b). A dominant paradigm preserves a pretrained LLM as the reasoning backbone, connects it to modality-specific encoders via projection modules, and performs multimodal instruction tuning to align new modalities with the language space (Liu et al., 2023b; Dai et al., 2023). This paradigm enables strong performance on multimodal tasks such as visual question answering and reasoning (Liu et al., 2024c; Yue et al., 2024; Chen et al., 2024).

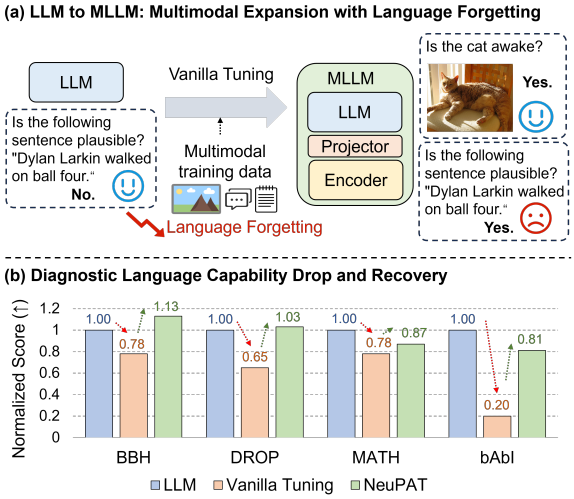

Figure 1: **Pure-text capability degradation and recovery after visual instruction tuning.** (a) Expanding an LLM into an MLLM with multimodal instruction data can induce language forgetting. (b) Vanilla tuning degrades performance on representative pure-text benchmarks, whereas NeuPAT recovers most of the lost capability. Scores are normalized by the original LLM performance on each benchmark.

Ideally, multimodal expansion should represent capability evolution: the model acquires new multimodal understanding while preserving the language intelligence accumulated during pretraining. This preservation is fundamental because modern MLLMs remain language-grounded systems, where language capabilities provide the foundation for reasoning, knowledge organization, and multimodal generalization.

However, this ideal preservation is not always

achieved. As shown in Figure 1, multimodal instruction tuning often causes substantial degradation on language reasoning benchmarks, with an average performance drop of approximately 39.8% across key language evaluations. This reveals that multimodal expansion is not purely additive; adapting the pretrained backbone toward multimodal distributions can interfere with existing language representations and overwrite capabilities essential for language-based reasoning. Therefore, an important question arises: *How can we expand the capability boundary of LLMs with new modalities while preserving the language intelligence that forms the foundation of multimodal intelligence?*

Recent studies have explored strategies to alleviate language capability degradation during multimodal expansion. Existing solutions mainly rely on external interventions, including text-only data replay (Lu et al., 2024; Bai et al., 2025), architectural modifications (Zhang et al., 2024; Wang et al., 2025a; Lu et al., 2025), and post-training model merging (Ratzlaff et al., 2024; Yu and Ananiadou, 2025; Wang et al., 2026; Li et al., 2025). Text replay retains pretrained abilities through additional language supervision but requires extra data and careful objective balancing. Architecture-based methods isolate multimodal adaptation from the language backbone at the cost of customized designs and added complexity. Model merging combines the original LLM and adapted MLLM after training, yet may introduce trade-offs between language preservation and multimodal adaptation. More importantly, these methods rely on auxiliary interventions rather than directly regulating backbone adaptation during multimodal learning.

In this work, we revisit multimodal expansion from the perspective of internal adaptation dynamics. We investigate whether the pretrained LLM contains heterogeneous adaptation capacities that can be selectively regulated during multimodal learning. Our key insight is that neurons within the pretrained backbone exhibit distinct adaptation patterns: some are more critical for preserving language intelligence, while others provide greater flexibility for absorbing new multimodal knowledge. Therefore, uniformly updating all neurons during multimodal tuning may unnecessarily disrupt language-critical representations while underutilizing the model's intrinsic plasticity.

Based on this insight, we propose **NeuPAT** (Neuron-aware Plasticity Allocation Tuning), a neuron-level capability-preserving tuning framework for multimodal expansion. NeuPAT introduces a lightweight probing stage to estimate neuron-wise adaptation characteristics using a small set of diagnostic samples, without introducing additional training data or modifying the model architecture. Based on these measurements, NeuPAT dynamically allocates update constraints during multimodal instruction tuning: neurons sensitive to language capability preservation are protected from excessive adaptation, while neurons with greater multimodal plasticity are encouraged to acquire new knowledge. By regulating internal adaptation dynamics, NeuPAT enables capability-preserving expansion from LLMs to MLLMs.

As shown in Figure 1, NeuPAT substantially reduces language capability degradation during multimodal expansion, recovering 90.0% of lost performance across 4 language reasoning benchmarks. Experiments across 6 LLMs, 11 language benchmarks, and 5 multimodal benchmarks demonstrate that NeuPAT consistently generalizes across model families and scales, providing an efficient and architecture-agnostic solution for preserving language intelligence during multimodal expansion.

Our contributions are summarized as follows:

- We reveal heterogeneous adaptation dynamics within LLM backbones during multimodal expansion, showing that different neurons exhibit distinct adaptation patterns associated with language preservation and multimodal learning.
- We propose NeuPAT, a neuron-aware plasticity allocation framework, enabling efficient and architecture-agnostic multimodal expansion while preserving language intelligence.
- Extensive experiments demonstrate that NeuPAT consistently preserves language intelligence across diverse LLM families and model scales while maintaining comparable multimodal performance, enabling scalable capability expansion from LLMs to MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

MLLMs typically adopt an encoder-connector-LLM architecture, mapping visual features into the language space via cross-attention, Q-Former, or lightweight projection modules (Alayrac et al.,

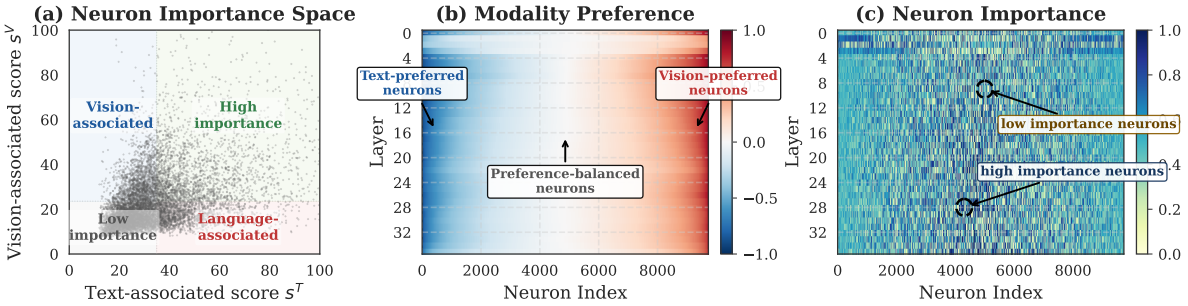

Figure 2: **Neuron modality preference and importance during multimodal expansion of LLMs.** (a) Raw text- and vision-associated importance scores, showing heterogeneous modality associations across neurons. (b) Layer-wise modality preference $P_{l,u} = v_{l,u} - t_{l,u}$, with neurons sorted by preference within each layer. (c) Overall importance $I_{l,u} = (v_{l,u} + t_{l,u})/2$ under the same ordering, showing diverse importance levels among neurons with similar preferences. These results reveal heterogeneous plasticity patterns in the pretrained backbone, motivating neuron-aware plasticity allocation during multimodal adaptation.

2022; Li et al., 2023a; Dai et al., 2023; Zhu et al., 2024; Liu et al., 2023b, 2024a). Recent models, including the Qwen-VL series, DeepSeek-VL, InternVL3, and LLaVA-OneVision, improve multimodal perception and reasoning with stronger encoders, larger corpora, and advanced post-training (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025; Lu et al., 2024; Wang et al., 2025b; Zhu et al., 2025; Li et al., 2024). Despite this progress, most MLLMs still adapt pretrained LLM backbones through image-text alignment and visual instruction tuning, potentially degrading their language capabilities and motivating our study.

2.2 Methods for Mitigating Forgetting

A common solution is to mix text-only and multimodal data (Lu et al., 2024; Bai et al., 2023; Wang et al., 2024; Bai et al., 2025), which increases training cost and requires careful ratio tuning. Existing alternatives apply general parameter-efficient or continual-learning methods (Hu et al., 2022; Kirkpatrick et al., 2017), modify the architecture (Zhang et al., 2024; Wang et al., 2025a; Lu et al., 2025), or merge the adapted model with the original LLM (Ratzlaff et al., 2024; Yu and Ananiadou, 2025; Wang et al., 2026; Li et al., 2025). In contrast, NeuPAT leverages neuron modality preferences to preserve language capabilities during multimodal learning without text replay, architectural changes, or post-hoc merging.

3 Neuron-Level Modality Preference Analysis

To enable capability-preserving multimodal expansion, we first analyze how different neurons within the pretrained LLM backbone respond to text and visual inputs. Our goal is to investigate whether the backbone contains heterogeneous plasticity patterns during multimodal adaptation. Without such knowledge, uniform update strategies, such as globally freezing or regularizing the backbone, may preserve language capabilities but unnecessarily restrict the model's ability to acquire new multimodal knowledge.

3.1 Modality-Associated Neuron Importance Estimation

For a Transformer layer l, we use u to index a neuron, corresponding to an intermediate hidden dimension together with its associated input- and output-side parameter slices. Let $h_{l,u}(x_t)$ denote the activation of neuron u at token position t, and let $W_{l,u}^{\text{out}}$ denote its output-side parameter slice that writes the neuron response back to the residual stream.

We construct lightweight text-only and visual probing sets, $\mathcal{D}_T$ and $\mathcal{D}_V$, respectively. Inspired by activation-aware importance estimation (Sun et al., 2024), we define the modality-associated importance score of neuron u under probing set $\mathcal{D}$ as

$$s_{l,u}(\mathcal{D}) = \left\|W_{l,u}^{\text{out}}\right\|_F \cdot \text{RMS}_{x\in\mathcal{D},\, t\in\mathcal{V}(x)} \left(\left\|h_{l,u}(x_t)\right\|_2\right), \quad (1)$$

where $\mathcal{V}(x)$ denotes the valid non-padding token positions. This score jointly considers neuron activation strength and its output contribution, providing a proxy for neuron importance under a specific input modality.

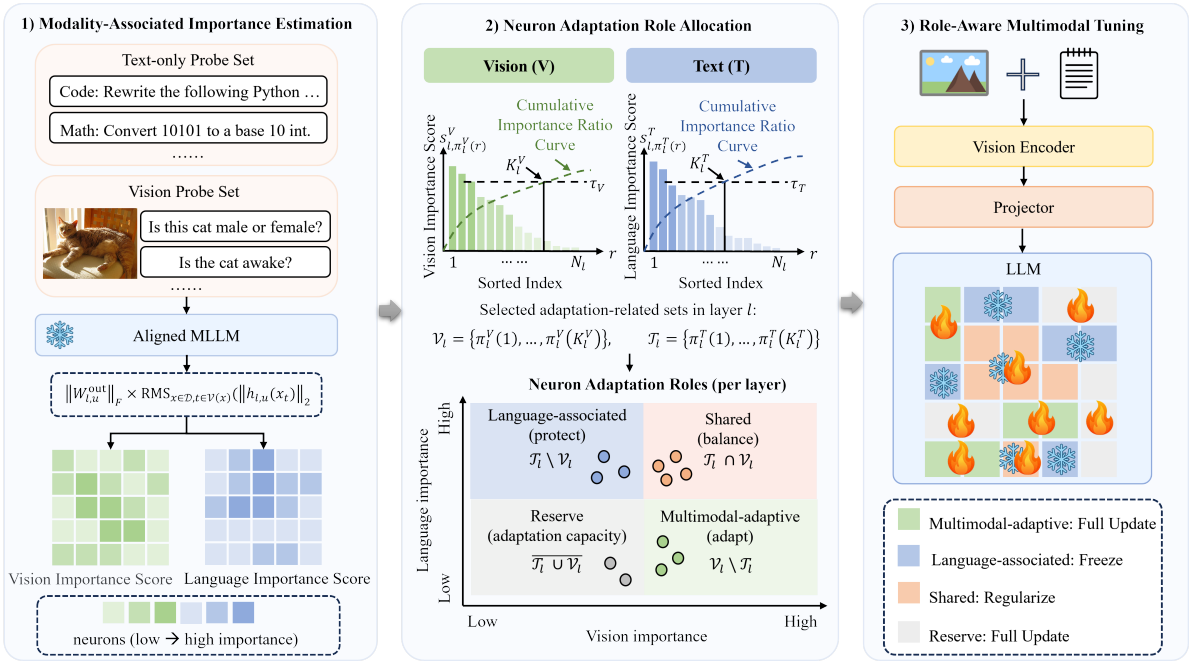

Figure 3: **Overview of NeuPAT**. NeuPAT identifies neuron adaptation roles through lightweight modality probing and assigns role-specific update constraints during multimodal tuning, preserving language intelligence while acquiring multimodal capabilities.

We compute the visual- and language-associated importance scores as

$$s^V_{l,u} = s_{l,u}(\mathcal{D}_V), \qquad s^T_{l,u} = s_{l,u}(\mathcal{D}_T). \quad (2)$$

For visualization and comparison within each layer, the scores are normalized as

$$v_{l,u} = \mathrm{Norm}(s^V_{l,u}), \qquad t_{l,u} = \mathrm{Norm}(s^T_{l,u}), \quad (3)$$

where $\mathrm{Norm}(\cdot)$ denotes layer-wise normalization.

Based on the normalized importance scores, we define two complementary properties:

$$P_{l,u} = v_{l,u} - t_{l,u}, \qquad I_{l,u} = \frac{v_{l,u} + t_{l,u}}{2}, \quad (4)$$

where $P_{l,u}$ measures the relative modality preference of a neuron, while $I_{l,u}$ measures its overall importance across both modalities. A positive $P_{l,u}$ indicates stronger association with visual inputs, whereas a negative value indicates stronger association with language inputs. Values close to zero indicate balanced importance across modalities.

3.2 Empirical Observations

Figure 2 illustrates neuron importance patterns from three perspectives. Figure 2(a) compares visual- and language-associated importance scores, revealing heterogeneous modality associations among neurons. Figures 2(b) and 2(c) further visualize the layer-wise modality preference $P_{l,u}$ and overall importance $I_{l,u}$, respectively, where neurons are sorted according to their preference within each layer.

Based on these analyses, we obtain three observations.

1) Heterogeneous modality-associated plasticity. Neurons exhibit diverse preferences toward language and visual inputs, indicating that the pre-trained backbone does not participate uniformly in multimodal adaptation. Consequently, applying identical updates to all neurons may unnecessarily perturb language-critical representations while limiting the utilization of neurons better suited for multimodal adaptation.

2) Different neurons require different update constraints. Neurons with balanced modality preferences are not homogeneous: some exhibit high importance under both input types, suggesting shared functionality across modalities, while others show relatively low importance and may provide additional flexibility for acquiring new multimodal knowledge. This indicates that different neurons require different levels of update constraint rather than uniform protection or adaptation.

3) Plasticity patterns vary across layers. The distributions of modality preference and impor-

tance vary substantially across Transformer layers, indicating that different layers exhibit distinct adaptation patterns. Therefore, fixed global allocation strategies may fail to capture layer-specific characteristics.

Together, these observations reveal heterogeneous adaptation requirements within the pretrained backbone: some neurons require protection to preserve language intelligence, while others provide flexibility for multimodal adaptation. This motivates neuron-aware allocation of update flexibility during multimodal expansion.

4 Methods

NeuPAT performs neuron-aware update allocation during multimodal instruction tuning. It estimates modality-associated importance, assigns neurons with different adaptation roles, and applies role-specific update constraints to balance language preservation and multimodal adaptation. Figure 3 illustrates the overall pipeline.

4.1 Modality-Associated Importance Estimation

Following the analysis in Section 3, NeuPAT first estimates modality-associated importance for each neuron using lightweight probing sets. Specifically, we construct visual and text-only probing sets and compute the visual- and language-associated importance scores, $s^V_{l,u}$ and $s^T_{l,u}$, using Eq. 1. These scores measure the contribution of individual neurons under different input modalities and provide the basis for subsequent adaptation role allocation.

For each layer l, the importance scores are normalized using layer-wise normalization using Eq. 2. The normalized scores are used to characterize the modality-associated importance distribution of neurons within each layer.

4.2 Neuron Adaptation Role Allocation

NeuPAT identifies neurons with different adaptation roles according to their modality-associated importance distributions. For each layer l, let $s^V_l \in \mathbb{R}^{N_l}$ and $s^T_l \in \mathbb{R}^{N_l}$ denote the visual- and language-associated importance scores of all N_l neurons.

For each modality $a \in \{V, T\}$, we select the smallest neuron subset whose cumulative importance accounts for a predefined coverage ratio τ_a:

$$K^a_l = \min \left\{ K : \frac{\sum_{r=1}^{K} s^a_{l,\pi^a_l(r)}}{\sum_{r=1}^{N_l} s^a_{l,\pi^a_l(r)} + \epsilon} \geq \tau_a \right\}, \tag{5}$$

where $\pi^a_l(\cdot)$ denotes the neuron indices ranked by modality-associated importance. The selected neuron sets for visual and language inputs are denoted as $\mathcal{V}_l$ and $\mathcal{T}_l$, respectively.

Based on the relationship between $\mathcal{V}_l$ and $\mathcal{T}_l$, we derive four neuron adaptation roles:

$$\begin{aligned} \mathcal{C}^{lang}_l &= \mathcal{T}_l \setminus \mathcal{V}_l, \quad & \mathcal{C}^{multi}_l &= \mathcal{V}_l \setminus \mathcal{T}_l, \\ \mathcal{C}^{shared}_l &= \mathcal{T}_l \cap \mathcal{V}_l, \quad & \mathcal{C}^{reserve}_l &= \overline{\mathcal{T}_l \cup \mathcal{V}_l}. \end{aligned} \tag{6}$$

where the complement in $\mathcal{C}^{reserve}_l$ is taken with respect to all neurons in layer l. Here, $\mathcal{C}^{lang}_l$ contains neurons primarily associated with language inputs and therefore requires protection during multimodal adaptation. $\mathcal{C}^{multi}_l$ contains neurons associated with visual inputs and provides adaptive capacity for acquiring multimodal knowledge. $\mathcal{C}^{shared}_l$ contains neurons important to both modalities and requires constrained updates, while $\mathcal{C}^{reserve}_l$ provides additional flexibility for multimodal learning.

4.3 Role-Aware Multimodal Tuning

Based on the assigned adaptation roles, NeuPAT regulates neuron-wise update flexibility during multimodal instruction tuning.

Language-associated neurons are frozen to preserve pretrained language representations. Multimodal-adaptive neurons receive full updates to absorb new multimodal knowledge. Reserve neurons are also fully optimized as additional adaptation capacity. Shared neurons, which participate in both language and multimodal processing, receive constrained updates to balance preservation and adaptation.

For shared neurons, we separate the corresponding parameters into input- and output-side components, denoted by $W^{\text{in}}_{l,u}$ and $W^{\text{out}}_{l,u}$. Their deviation from pretrained parameters is constrained by:

$$\begin{aligned} \mathcal{R}_{\text{shared}} = \sum_{l} \sum_{u \in \mathcal{C}^{\text{shared}}_l} \Bigg[& \lambda_{\text{in}} \left\| W^{\text{in}}_{l,u} - W^{\text{in},0}_{l,u} \right\|^2_F \\ & + \lambda_{\text{out}} \left(1 - \cos(W^{\text{out}}_{l,u}, W^{\text{out},0}_{l,u}) \right) \Bigg]. \end{aligned} \tag{7}$$

The input-side constraint limits changes to neuron activation behavior, while the output-side cosine

Table 1: Comparison on 11 language benchmarks grouped into Language and Logical Reasoning Tasks, Math and Code Reasoning Tasks, and General Tasks. $\Delta(2)-(1)$ denotes Vanilla Tuning − LLM, while $\Delta(9)-(2)$ denotes NeuPAT − Vanilla Tuning. Bold values indicate the best results among multimodally adapted models.

Method	Language and Logical Reasoning Tasks					Math and Code Reasoning Tasks				General Tasks					Overall
	BBH	bAbI	DROP	LogiQA2	Avg.	MATH-500	MBPP	GSM8K	Avg.	SocialIQA	CoQA	GPQA	ARC-C	Avg.	Avg.
(1) LLM	47.83	13.20	15.30	35.94	28.07	64.60	65.80	86.20	72.20	50.10	67.53	36.38	55.97	52.50	48.99
(2) + Vanilla Tuning	37.12	2.68	9.90	32.95	20.66	50.40	62.60	83.47	65.49	46.78	67.37	34.15	55.72	51.01	43.92
$\Delta(2)-(1)$	-10.71	-10.52	-5.40	-2.99	**-7.41**	-14.20	-3.20	-2.73	**-6.71**	-3.32	-0.16	-2.23	-0.25	**-1.49**	**-5.07**
(3) + LoRA	49.72	1.57	11.42	33.97	24.17	46.20	65.40	81.96	64.52	**51.38**	**70.58**	36.16	55.63	53.44	45.82
(4) + EWC	42.50	2.74	12.90	35.88	23.51	52.80	66.00	85.52	68.11	47.65	69.77	35.71	**57.51**	52.66	46.27
(5) + WINGS	51.76	2.70	11.03	34.16	24.91	53.20	63.20	79.91	65.44	50.00	70.50	35.49	56.48	53.12	46.22
(6) + TIES	38.69	2.67	14.70	35.62	22.92	51.40	63.20	84.15	66.25	46.93	67.72	34.60	56.06	51.33	45.07
(7) + L2M	41.38	2.67	14.47	35.81	23.58	50.00	63.80	84.00	65.93	46.52	68.52	34.38	55.97	51.35	45.23
(8) + PlaM	36.37	2.67	14.61	35.69	22.34	50.80	62.80	83.32	65.64	46.47	67.45	34.60	55.89	51.10	44.61
(9) + **NeuPAT**	**53.88**	**10.67**	**15.74**	**36.39**	**29.17**	**56.40**	**66.60**	**85.97**	**69.66**	50.26	69.53	**37.27**	56.91	**53.49**	**49.06**
$\Delta(9)-(2)$	+16.76	+7.99	+5.84	+3.44	**+8.51**	+6.00	+4.00	+2.50	**+4.17**	+3.48	+2.16	+3.12	+1.19	**+2.49**	**+5.14**

Table 2: Comparison on multimodal benchmarks. Δ denotes NeuPAT − Vanilla Tuning.

Method	MMB	RWQA	MMMU	POPE	MMS	Avg.
Vanilla Tuning	72.08	56.60	43.11	87.69	45.40	60.98
LoRA	68.21	53.59	39.44	86.57	42.56	58.07
EWC	67.53	50.59	43.78	84.99	40.67	57.51
WINGS	73.88	57.78	43.44	87.86	47.91	62.17
TIES	72.42	55.56	43.00	87.74	45.81	60.91
L2M	72.94	55.16	42.89	87.67	46.24	60.98
PlaM	72.77	56.47	43.00	87.77	45.10	61.02
NeuPAT	72.48	56.99	42.22	88.72	44.85	61.05
Δ	+0.40	+0.39	-0.89	+1.03	-0.55	+0.07

constraint preserves the direction of neuron contributions to the residual stream. During optimization, gradients of language-associated neurons are masked, while other neuron roles follow their assigned update constraints.

The final training objective is $\mathcal{L} = \mathcal{L}_{\text{ori}} + \mathcal{R}_{\text{shared}}$,where $\mathcal{L}_{\text{ori}}$ denotes the original autoregressive language-modeling objective.

5 Experiments

5.1 Experimental Settings

5.1.1 Implementation Details

We use RICE-ViT-Large-Patch14-560 as the vision tower (Xie et al., 2025) and Qwen3-4B-Instruct-2507 as the default language backbone (Yang et al., 2025). Generalization is evaluated on five additional backbones: Qwen3-0.6B, Phi-4-Mini-Instruct, Qwen2.5-7B-Instruct, Llama3.1-8B-Instruct, and Qwen2.5-14B-Instruct (Yang et al., 2025; Abouelenin et al., 2025; Yang et al., 2024; Grattafiori et al., 2024). Training consists of image-text alignment on LLaVA-558K (Liu et al., 2024a) with only the multimodal adapter updated, followed by NeuPAT-based visual instruction tuning on LLaVA-NeXT-780K (Liu et al., 2024b). We set $\tau_a = 0.8$ and $\lambda_{\text{in}} = \lambda_{\text{out}} = 0.1$. All models are trained with Adam on 8 NVIDIA A100 GPUs.

5.1.2 Baselines

We compare NeuPAT with three baseline groups: **(1) reference models,** including the original LLM and Vanilla Tuning; **(2) general adaptation methods,** including LoRA (Hu et al., 2022) and EWC (Kirkpatrick et al., 2017); and **(3) MLLM-specific preservation methods,** including the architecture-based WINGS (Zhang et al., 2024) and post-hoc merging approaches such as TIES (Ratzlaff et al., 2024), Locate-then-Merge (Yu and Ananiadou, 2025), and PlaM (Wang et al., 2026). The original LLM provides a reference for language capability, while Vanilla Tuning reflects the forgetting caused by standard visual instruction tuning.

5.1.3 Evaluation Protocol

We evaluate language and multimodal capabilities using `lm-evaluation-harness` (Gao et al., 2024) and `lmms-eval` (Zhang et al., 2025), respectively. The language suite contains 11 benchmarks covering knowledge, mathematics, logic, coding, and reading comprehension: SocialIQA (Sap et al., 2019), ARC-Challenge (Clark et al., 2018), GPQA (Rein et al., 2023), MATH-500 (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), LogiQA2 (Liu et al., 2023a), BBH (Suzgun et al., 2023), bAbI (Weston et al., 2015), MBPP (Austin et al., 2021), DROP (Dua et al., 2019) and CoQA (Reddy et al., 2019). Multimodal performance is evaluated on MMBench-EN (Liu et al., 2024c), RealWorldQA (xAI, 2024), POPE (Li et al., 2023c), MMMU (Yue et al., 2024) and MM-Star (Chen et al., 2024).

Table 3: Generalization across different LLM backbones. For each backbone, we report the original LLM, Vanilla Tuning, and our method. We show representative language benchmarks and multimodal benchmarks. Δ denotes NeuPAT $-$ Vanilla Tuning.

Backbone	Method	Language Benchmarks						Multimodal Benchmarks					
		SIQA	GSM8K	BBH	MBPP	CoQA	Avg.	MMB	RWQA	MMMU	POPE	MMStar	Avg.
Qwen3-0.6B	LLM	40.38	40.56	33.39	27.40	57.57	39.86	–	–	–	–	–	–
	+ Vanilla Tuning	39.00	35.18	28.69	22.60	55.93	36.28	52.58	45.10	30.67	85.42	37.23	50.20
	+ NeuPAT	40.33	40.11	33.44	28.20	59.42	40.30	52.52	44.75	32.67	84.26	37.07	50.25
	Δ	+1.33	+4.93	+4.75	+5.60	+3.49	+4.02	-0.06	-0.35	+2.00	-1.16	-0.16	+0.05
Phi-4-Mini-Instruct	LLM	49.59	83.62	52.82	55.40	78.40	63.97	–	–	–	–	–	–
	+ Vanilla Tuning	46.11	73.92	40.04	47.60	65.23	54.58	50.52	45.10	36.78	77.16	30.58	48.03
	+ NeuPAT	50.06	81.45	51.29	53.80	80.04	63.33	50.89	44.31	36.22	77.82	31.26	48.10
	Δ	+3.95	+7.53	+11.25	+6.20	+14.81	+8.75	+0.37	-0.79	-0.56	+0.66	+0.68	+0.07
Qwen2.5-7B-Instruct	LLM	51.59	76.50	45.81	47.60	78.74	60.05	–	–	–	–	–	–
	+ Vanilla Tuning	50.20	74.83	41.98	40.20	73.63	56.17	64.78	52.03	41.00	87.12	40.56	57.10
	+ NeuPAT	55.89	78.70	47.53	42.80	78.80	60.74	64.64	52.01	42.56	86.54	40.37	57.22
	Δ	+5.69	+3.87	+5.55	+2.60	+5.17	+4.57	-0.14	-0.02	+1.56	-0.58	-0.19	+0.12
Llama3.1-8B-Instruct	LLM	49.85	78.17	44.59	58.40	77.83	61.77	–	–	–	–	–	–
	+ Vanilla Tuning	49.74	67.55	34.91	54.20	77.50	56.78	60.14	26.67	35.89	82.23	36.54	48.29
	+ NeuPAT	49.95	74.55	41.68	56.80	80.48	60.69	61.94	44.97	36.22	80.90	33.94	51.59
	Δ	+0.21	+7.00	+6.77	+2.60	+2.98	+3.91	+1.80	+18.30	+0.33	-1.33	-2.60	+3.30
Qwen2.5-14B-Instruct	LLM	54.04	79.83	52.54	66.80	78.20	66.28	–	–	–	–	–	–
	+ Vanilla Tuning	51.23	79.30	39.87	65.60	74.44	62.09	75.17	56.34	47.67	87.27	50.26	63.34
	+ NeuPAT	55.22	85.44	52.59	66.80	77.51	67.51	74.31	55.64	48.11	87.67	50.14	63.17
	Δ	+3.99	+6.14	+12.72	+1.20	+3.07	+5.42	-0.86	-0.70	+0.44	+0.40	-0.12	-0.17

5.2 Experimental Results

5.2.1 Main Results

Tables 1 and 2 report results on 11 language and 5 multimodal benchmarks. Vanilla Tuning lowers the language average from 48.99 to 43.92, with drops exceeding 10 points on BBH, bAbI, and MATH-500. NeuPAT recovers 5.14 points, reaching 49.06 and slightly surpassing the original LLM, with particularly strong gains on language and logical reasoning tasks. Meanwhile, it maintains comparable multimodal performance, improving the average from 60.98 to 61.05. Although WINGS achieves the highest multimodal average, NeuPAT provides the best overall language performance and a stronger balance between language preservation and multimodal adaptation.

5.2.2 Generalization across LLM Backbones

To evaluate cross-backbone generalization, we replace the default LLM with Qwen3-0.6B, Phi-4-Mini-Instruct, Qwen2.5-7B/14B-Instruct, and Llama3.1-8B-Instruct. Due to space limitations, Table 3 reports five representative language benchmarks, while the complete results are provided in the appendix. Vanilla Tuning consistently degrades language performance across all tested backbones, whereas NeuPAT improves the reported text average over Vanilla Tuning by 4.02, 8.75, 4.57, 3.91, and 5.42 points, respectively, while maintaining comparable multimodal performance. For Qwen3-0.6B and both Qwen2.5 backbones, NeuPAT even surpasses the original LLM average. These consistent improvements across different model families and scales demonstrate that NeuPAT is not tied to a specific language backbone.

5.2.3 Ablation Studies

Global vs. Neuron-Aware Update. We compare NeuPAT with three global update schemes. *Freeze* freezes all neurons, *Update* uniformly updates the backbone, and *Reg.* applies the shared-neuron regularization globally. As shown in Table 5, global strategies suffer from either limited multimodal adaptation or insufficient language preservation. NeuPAT achieves the best balance, obtaining language and multimodal averages of 54.41 and 72.73, demonstrating the effectiveness of neuron-wise update allocation.

Neuron Allocation Strategy. We compare importance-guided role allocation with *Random* and *Fixed-ratio* baselines. *Random* preserves category sizes but randomly assigns neuron roles, while *Fixed-ratio* applies the same selection ratio across layers. As shown in Table 6, NeuPAT consistently outperforms both baselines, validating the importance of reliable neuron identification and layer-adaptive allocation.

Neuron-Wise Update Constraint. We further evaluate each role-specific update strategy by modifying one constraint at a time. As shown in Table 4, updating language-associated neurons decreases text performance, while freezing vision-associated

Table 4: Ablation of neuron-wise adaptation role allocation. ✓: full update, ×: freeze, Reg.: regularized update.

Variants	Neuron Adaptation Role				Language Benchmarks				Multimodal Benchmarks			
	Lang.	Multi.	Shared	Reserve	GSM8K	MBPP	bAbI	Avg.	MMB	RWQA	POPE	Avg.
w/o Language Freeze	✓	✓	Reg.	✓	83.90	64.80	5.87	51.52	72.11	57.30	87.20	72.20
w/o Multimodal Update	×	×	Reg.	✓	84.44	67.10	11.85	54.46	69.69	55.16	86.61	70.49
w/o Reserve Update	×	✓	Reg.	×	84.76	66.80	11.64	54.40	71.05	56.34	86.94	71.44
Shared Full Update	×	✓	✓	✓	83.61	64.00	9.94	52.52	72.34	57.25	87.27	72.29
Shared Freeze	×	✓	×	✓	85.06	67.20	15.16	55.81	69.93	55.03	87.30	70.75
Ours	×	✓	Reg.	✓	85.97	66.60	10.67	54.41	72.48	56.99	88.72	72.73

Table 5: Ablation of global update schemes and role-aware update allocation.

Update strategy	Language Benchmarks				Multimodal Benchmarks			
	GSM8K	MBPP	bAbI	Avg.	MMB	RWQA	POPE	Avg.
Freeze	86.05	67.10	13.65	55.60	68.64	53.99	86.03	69.55
Update	83.47	62.60	2.68	49.58	72.08	56.60	87.69	72.12
Reg.	84.99	65.50	9.33	53.27	72.08	54.25	86.03	70.79
Ours	85.97	66.60	10.67	54.41	72.48	56.99	88.72	72.73

Table 6: Ablation of neuron role allocation strategies.

Allocation strategy	Language Benchmarks				Multimodal Benchmarks			
	GSM8K	MBPP	bAbI	Avg.	MMB	RWQA	POPE	Avg.
Random	83.69	65.40	1.94	50.34	69.76	54.51	87.52	70.60
Fixed-ratio	84.90	66.20	10.24	53.78	72.16	57.25	87.28	72.23
Ours	85.97	66.60	10.67	54.41	72.48	56.99	88.72	72.73

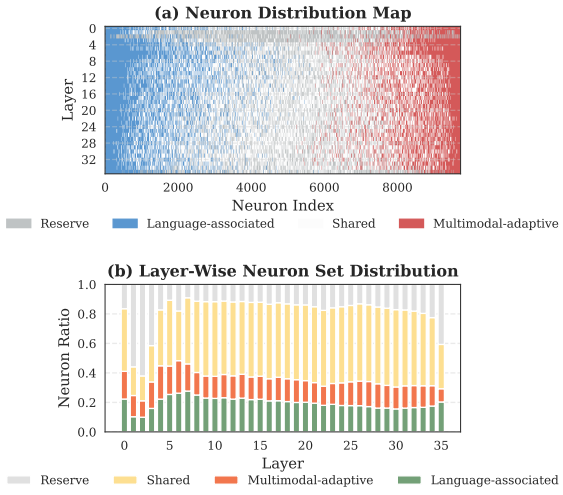

Figure 4: **Layer-wise adaptation role distribution.** (a) Adaptation roles identified by importance-guided allocation. (b) Layer-wise proportions of neuron roles.

or reserve neurons limits multimodal adaptation. For shared neurons, full updating harms language preservation and full freezing sacrifices multimodal performance. These results support the proposed strategy of freezing language neurons, updating multimodal and reserve neurons, and regularizing shared neurons.

5.2.4 Neuron Distribution Visualization

We visualize the assigned neuron roles and their layer-wise distributions in Figure 4. As shown in Figure 4(a), the resulting neuron map is closely consistent with the modality-associated response patterns observed in Figure 2. Language-associated neurons are concentrated in text-preferred regions, whereas multimodal-adaptive neurons primarily occupy vision-preferred regions. In contrast, shared and reserve neurons are located mainly in balanced-response regions, but differ markedly in their overall response strength: shared neurons respond strongly to both input types, while reserve neurons remain weakly engaged and may provide underutilized capacity for multimodal adaptation. Figure 4(b) further shows that shared neurons constitute the largest group in most layers, accounting for approximately 45%, while language-associated and multimodal-adaptive neurons each represent around 20%, and reserve neurons account for roughly 15%. These consistent yet layer-dependent distributions reveal substantial heterogeneity in neuron functionality and adaptation capacity, supporting the need for layer-adaptive neuron allocation and differentiated plasticity control within the LLM backbone.

6 Conclusion

Multimodal expansion enables LLMs to acquire capabilities beyond language, but may compromise the language intelligence inherited from pretraining. We show that this degradation stems from heterogeneous adaptation behaviors within the pretrained LLM backbone and introduce NeuPAT, a neuron-aware update allocation framework for capability-preserving multimodal tuning. By selectively regulating neuron update flexibility, NeuPAT balances language preservation and multimodal adaptation without additional data, architectural changes, or post-training merging. Experiments across diverse LLM backbones demonstrate that NeuPAT is an efficient, architecture-agnostic solution for reliable LLM-to-MLLM expansion.

7 Limitations

NeuPAT has several limitations. First, although we evaluate it across diverse LLM families and model scales, its effectiveness on substantially larger backbones remains unverified. Larger models may exhibit more distributed and complex adaptation patterns, and the current neuron-wise role allocation mechanism may require further investigation at greater scale. Second, our experiments focus on vision-language expansion. Extending NeuPAT to other modalities, such as audio, video, embodied interaction, or unified multimodal systems, may introduce different alignment dynamics and modality interactions, potentially requiring modality-specific importance estimation and update strategies. Finally, this work considers a single-stage multimodal expansion process. In continual or sequential learning scenarios, newly introduced modalities and tasks may interact with both pretrained language capabilities and previously acquired multimodal knowledge. Developing mechanisms for stable long-term capability accumulation without progressive interference remains an important direction for future research.

8 Ethical Considerations

All training data and evaluation experiments in this work are based on publicly available datasets and benchmarks. And we emphasize that the proposed neuron allocation strategies are response-based functional approximations rather than causal explanations of model behavior. Models trained with NeuPAT should therefore undergo standard safety, fairness, and robustness evaluations before deployment, especially in high-stakes applications. AI tools were used only for language polishing. All research ideas, experiments, analyses, and manuscript organization were completed by the authors.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

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*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Tingchao Fu, Wenkai Wang, Fanxiao Li, Huadong Zhang, Jinhong Zhang, Dayang Li, Yunyun Dong, Renyang Liu, and Wei Zhou. 2026. Correct when paired, wrong when split: Decoupling and editing modality-specific neurons in mllms. *arXiv preprint arXiv:2606.17057*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Bo Li, Ningyuan Deng, Tianyu Dong, Shaobo Wang, Shaolin Zhu, and Lijie Wen. 2026. Mnaft: modality neuron-aware fine-tuning of multimodal large language models for image translation. *Science China Information Sciences*, 69(5).

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Zhihua Li, Haoguang Wen, Wenpeng Hu, Zhunchen Luo, Lingqiang Chen, Sijia Wang, and Shuyi Wu. 2025. Dpimerge: An efficient dynamic parameter interpolation framework for alleviating pure text forgetting in multimodal large models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 3–15. Springer.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. 2025. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5913–5933.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Xudong Lu, Yinghao Chen, Renshou Wu, Haohao Gao, Xi Chen, Xue Yang, Xiangyu Zhao, Aojun Zhou, Fangyuan Li, Yafei Wen, et al. 2025. Genieblue: Integrating both linguistic and multimodal capabilities for large language models on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4198–4210.

Zixuan Qin, Qingchen Yu, Kunlin Lyu, Zhaoxin Fan, and Yifan Sun. 2025. The achilles' heel of llms: How altering a handful of neurons can cripple language abilities. *arXiv preprint arXiv:2510.10238*.

Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. 2024. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. *arXiv preprint arXiv:2412.03467*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4463–4473.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bin Wang, Chunyu Xie, Dawei Leng, and Yuhui Yin. 2025a. Iaa: Inner-adaptor architecture empowers frozen large language model with multimodal capabilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21035–21043.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zijing Wang, Yongkang Liu, Mingyang Wang, Ercong Nie, Deyuan Chen, Zhengjie Zhao, Shi Feng, Daling Wang, Xiaocui Yang, Yifei Zhang, et al. 2026. Plam: Training-free plateau-guided model merging for better visual grounding in mllms. *arXiv preprint arXiv:2601.07645*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

xAI. 2024. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v. Introduces the RealWorldQA benchmark.

Yin Xie, Kaicheng Yang, Xiang An, Kun Wu, Yongle Zhao, Weimo Deng, Zimin Ran, Yumeng Wang, Ziyong Feng, Roy Miles, et al. 2025. Region-based cluster discrimination for visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1793–1803.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Zeping Yu and Sophia Ananiadou. 2025. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal llms. *arXiv preprint arXiv:2505.16703*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal llms without text-only forgetting. *Advances in Neural Information Processing Systems*, 37:31828–31853.

Xiutian Zhao, Björn Schuller, and Berrak Sisman. 2026. Discovering and causally validating emotion-sensitive neurons in large audio-language models. *arXiv preprint arXiv:2601.03115*.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Algorithm 1 NeuPAT Training Procedure

Require: Aligned model Θ^0, probing sets $\mathcal{D}_T, \mathcal{D}_V$, training set $\mathcal{D}_{\text{ori}}$, thresholds τ_T, τ_V
Ensure: Tuned model Θ

Compute text and vision response scores $\{s^T_{l,u}, s^V_{l,u}\}$ using $\mathcal{D}_T$ and $\mathcal{D}_V$
for each layer l **do**
    Select important sets $\mathcal{T}_l$ and $\mathcal{V}_l$ using cumulative importance thresholds τ_T and τ_V
    $\mathcal{C}_l^{text} \leftarrow \mathcal{T}_l \setminus \mathcal{V}_l$
    $\mathcal{C}_l^{vision} \leftarrow \mathcal{V}_l \setminus \mathcal{T}_l$
    $\mathcal{C}_l^{high} \leftarrow \mathcal{T}_l \cap \mathcal{V}_l$
    $\mathcal{C}_l^{low} \leftarrow \overline{\mathcal{T}_l \cup \mathcal{V}_l}$
end for
$\Theta \leftarrow \Theta^0$
for each minibatch $\mathcal{B} \subset \mathcal{D}_{\text{ori}}$ **do**
    Compute $\mathcal{L}_{\text{ori}}$ and $\mathcal{L} = \mathcal{L}_{\text{ori}} + \mathcal{R}_{\text{high}}$
    Freeze text neurons, update vision and low response neurons with $\mathcal{L}_{\text{ori}}$, and update high response neurons with $\mathcal{L}$
end for
return Θ

A Method Details

A.1 Overall Algorithm

The complete pseudocode of our NeuPAT method is shown in Algorithm 1.

B Additional Related Work

B.1 Neuron-Level Analysis

Recent studies have investigated the functional specialization of individual neurons in language and multimodal models. Fu et al. (Fu et al., 2026) reveal that knowledge in MLLMs can be distributed across decoupled modality-specific neuron pathways and localize these neurons to improve knowledge editing under different modality inputs. Zhao et al. (Zhao et al., 2026) identify emotion-sensitive neurons in large audio-language models and causally validate their roles through neuron suppression and activation steering. For multimodal unlearning, Liu et al. (Liu et al., 2025) locate neurons according to their relative importance across modalities and selectively prune them to remove targeted knowledge while preserving general model utility. MNAFT (Li et al., 2026) identifies modality-relevant neurons using activation and gradient information and selectively fine-tunes them for image translation to reduce interference across

Table 7: Training configurations for the two-stage pipeline.

Configuration	Stage 1: Alignment	Stage 2: Instruction Tuning
Dataset	LLaVA-558K	LLaVA-NeXT-780K
Training steps	2,500	3,500
Trainable modules	Adapter only	LLM, adapter, and vision encoder
Global batch size	8	224
Micro-batch size	1	1
Gradient accumulation	1	28
Number of GPUs	8	8
Peak learning rate	1×10^{-4}	1×10^{-5}
Minimum learning rate	1×10^{-6}	1×10^{-6}
Warmup ratio	0.002	0.002
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-5}$)	
LR scheduler	Cosine decay	
Weight decay	0	
Gradient clipping	1.0	
Precision	BF16	
Sequence length	32,768	
Tensor / pipeline parallelism	1 / 1	
Image resolution	Default	1,000
Offline packing	Enabled	Disabled

languages and modalities. Qin et al. (Qin et al., 2025) further show that LLMs contain ultra-sparse critical neurons whose perturbation can severely impair overall language ability.

These studies associate neuron-level structures with modality-specific knowledge, task behaviors, and fundamental model capabilities, but primarily focus on knowledge editing, emotion control, unlearning, task-specific image translation, or vulnerability analysis. In contrast, we study neuron-wise adaptation dynamics during general multimodal learning. NeuPAT uses lightweight text and vision probing to estimate heterogeneous neuron responses, followed by neuron adaptation role allocation and role-aware multimodal tuning. By protecting language-sensitive neurons, promoting adaptation through vision-responsive and underutilized neurons, and regularizing neurons important to both text and vision, NeuPAT preserves pretrained language capabilities while supporting multimodal learning.

C Detailed Experimental Setup

C.1 Probing Sets

We construct separate text-only and vision probing sets to estimate modality-specific neuron responses. The vision probing set contains $N_V = 2048$ samples randomly drawn from LLaVA-NeXT-780K (Liu et al., 2024b). The text probing set contains $N_T = 2048$ prompts sampled from publicly available general-domain text datasets that are disjoint from all evaluation benchmarks. To improve coverage, we include prompts from instruction-following, factuality, and general reasoning domains, including CodeAlpaca-20k (Chaudhary, 2023), MetaMathQA (Yu et al., 2023), databricks-dolly-15k (Conover et al., 2023) and HaluEval (Li et al., 2023b). We allocate samples evenly across data sources. When a source contains fewer samples than its assigned quota, the remaining quota is redistributed among the other sources.

Both probing sets are used solely for forward-pass activation statistics. They do not contribute to the training objective or parameter updates. We use the same processor and chat-template pipeline for both sets, while providing images only for the vision probing samples.

C.2 Training Details

As shown in Table 7, we follow a two-stage training pipeline. In Stage 1, we perform image-text alignment on LLaVA-558K for 2,500 steps, where only the multimodal adapter is optimized. We use a global batch size of 8 on 8 A100 GPUs, with a micro-batch size of 1. The peak learning rate is 1×10^{-4} and is decayed to 1×10^{-6} using a cosine schedule. The warmup ratio is set to 0.002, corresponding to approximately 5 warmup steps.

In Stage 2, we conduct visual instruction tuning on LLaVA-NeXT-780K for 3,500 steps. The language model, multimodal adapter, and vision encoder are included in optimization, while NeuPAT applies plasticity-guided multimodal tuning strategies to the neurons in the language backbone. Training is performed on 8 A100 GPUs with a global batch size of 224, a micro-batch size of 1, and 28 gradient-accumulation steps. The peak learning rate is 1×10^{-5}, with the same minimum learning rate of 1×10^{-6} and a warmup ratio of 0.002. Both stages use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-5}$, together with cosine learning-rate decay, zero weight decay, and gradient clipping at 1.0.

C.3 Baseline Implementations

LoRA. LoRA (Hu et al., 2022) inserts trainable low-rank adapters into selected linear layers while freezing the original parameters. We set the rank to $r = 32$, the scaling factor to $\alpha = 64$, and the learning rate to 1×10^{-4}. The model is trained on the same multimodal instruction data as Vanilla

Table 8: Sensitivity analysis of the target importance mass τ_a on language and multimodal benchmarks. The selected threshold $\tau_a = 0.8$ is highlighted.

τ_a	Language Benchmarks												Multimodal Benchmarks					
	SIQA	ARC-C	GPQA	MATH	GSM8K	LogiQA2	BBH	bAbI	MBPP	DROP	CoQA	Avg.	MMB	RWQA	MMMU	POPE	MMStar	Avg.
0.6	49.54	56.57	35.27	54.40	84.84	36.77	53.46	5.87	65.40	14.57	69.12	47.80	73.28	57.65	43.22	87.59	46.20	61.59
0.7	50.51	56.40	36.16	55.60	85.14	35.37	53.17	11.88	66.60	14.37	69.85	48.64	72.51	57.52	43.67	87.04	45.59	61.27
0.8	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
0.9	49.69	55.97	35.71	55.40	86.43	35.69	53.51	13.99	67.80	15.11	70.63	49.08	71.39	54.77	43.11	86.73	43.47	59.89

Table 9: Sensitivity analysis of the size of probing set on language and multimodal benchmarks. The selected size 2048 is highlighted.

N_a	Language Benchmarks												Multimodal Benchmarks					
	SIQA	ARC-C	GPQA	MATH	GSM8K	LogiQA2	BBH	bAbI	MBPP	DROP	CoQA	Avg.	MMB	RWQA	MMMU	POPE	MMStar	Avg.
512	50.05	57.08	35.94	53.40	85.97	36.20	52.85	10.41	67.00	15.18	70.45	48.59	72.39	57.25	41.33	87.00	46.17	60.83
1024	50.36	56.83	35.94	52.00	84.76	36.07	54.12	14.89	67.00	14.17	70.32	48.77	71.99	56.73	43.11	87.28	44.47	60.72
2048	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
4096	50.56	56.40	35.94	52.22	85.44	37.15	53.39	11.16	67.20	13.89	70.62	48.54	70.88	56.99	42.56	87.13	45.89	60.69

Tuning.

EWC. EWC (Kirkpatrick et al., 2017) estimates parameter importance using the diagonal Fisher information and penalizes changes to important parameters. We estimate the Fisher information at the shared Stage-1 checkpoint using 2,048 text-only instruction samples that are disjoint from all evaluation benchmarks. The Fisher statistics are averaged over samples, and the EWC penalty is applied only to the language-model parameters. We search λ_{EWC} over $\{0.1, 1, 10, 100\}$ and select $\lambda_{\text{EWC}} = 10$ according to the text-multimodal performance. All remaining training settings are identical to Vanilla Tuning.

WINGS. WINGS (Zhang et al., 2024) introduces parallel visual and textual learners into attention layers to reduce over-reliance on visual tokens. Their outputs are fused with the original attention output through a learned router. We reproduce WINGS using the architecture and hyperparameter settings recommended in the original paper.

TIES. TIES (Ratzlaff et al., 2024) is a training-free merging method that sparsifies the task vector between the visually tuned model and the original LLM. We follow the original paper for the task-vector density, merge coefficient, and merging procedure.

Locate-then-Merge. Locate-then-Merge (Yu and Ananiadou, 2025) identifies high-impact neurons from parameter changes, suppresses low-impact updates, and restores selected neurons through replacement or rescaling. We follow the original paper for the neuron-retention ratio, parameter-sparsification ratio, restoration strategy, and associated coefficients. The hyperparameters are selected from the recommended ranges reported by the authors.

PlaM. PlaM (Wang et al., 2026) locates a plateau layer through layer-wise vision-token masking and merges subsequent layers with the original LLM. We follow the original paper to determine the plateau layer and search the merge coefficients within its recommended range. Earlier visual-alignment layers remain unchanged, while the selected later layers are linearly merged with the original language backbone.

D Additional Experimental Results

D.1 Sensitivity Analysis

All sensitivity analyses are conducted after fixing the default configuration used in the main experiments. We vary one hyperparameter at a time while keeping all other settings unchanged. These experiments are intended to evaluate robustness.

Target Importance Mass τ_a. We analyze the sensitivity to the target importance mass τ_a used in neuron adaptation role allocation. As shown in Table 8, language performance generally improves as τ_a increases. A larger τ_a retains more cumulative response mass for each modality, thereby expanding the selected neuron sets and reducing the proportion of low response neurons. This tends to assign more neurons to text-related or high response neurons, strengthening language preservation but leaving less unconstrained capacity for

Table 10: Sensitivity analysis of the high response neuron regularization coefficients λ_{in} and λ_{out}. The selected setting $\lambda_{\text{in}} = \lambda_{\text{out}} = 0.1$ is highlighted.

$\lambda_{\text{in}} = \lambda_{\text{out}}$	Language Benchmarks												Multimodal Benchmarks					
	SIQA	ARC-C	GPQA	MATH	GSM8K	LogiQA2	BBH	bAbI	MBPP	DROP	CoQA	Avg.	MMB	RWQA	MMMU	POPE	MMStar	Avg.
0.05	50.26	57.42	36.61	52.60	85.97	34.41	52.02	10.09	67.00	15.58	69.70	48.33	72.68	57.91	42.11	87.19	46.29	61.24
0.1	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
0.5	50.31	56.57	37.28	54.20	84.08	36.26	52.79	13.86	68.00	14.75	70.78	48.99	71.65	56.21	43.00	87.41	45.82	60.82
1.0	50.41	56.14	37.05	55.60	85.90	36.26	52.30	13.38	66.60	13.71	70.85	48.93	71.65	55.95	43.22	87.41	44.87	60.62

Table 11: Complete cross-backbone results on 11 language and 5 multimodal benchmarks. "LLM" denotes the original language backbone before multimodal training.

Task	Qwen3-0.6B			Phi-4-Mini-Instruct			Qwen2.5-7B-Instruct			Llama3.1-8B-Instruct			Qwen2.5-14B-Instruct		
	LLM	Vanilla Tuning	NeuPAT	LLM	Vanilla Tuning	NeuPAT	LLM	Vanilla Tuning	NeuPAT	LLM	Vanilla Tuning	NeuPAT	LLM	Vanilla Tuning	NeuPAT
Language Benchmarks															
MATH-500	14.00	13.80	13.40	39.20	32.00	34.00	44.40	28.00	35.20	36.80	28.80	35.80	43.40	25.60	39.00
BBH	33.39	28.69	33.44	52.82	40.04	51.29	45.81	41.98	47.53	44.59	34.91	41.68	52.54	39.87	52.59
bAbI	2.50	1.93	2.21	2.65	1.06	2.78	2.71	0.60	2.56	0.00	0.14	0.81	1.17	0.03	1.48
MBPP	27.40	22.60	28.20	55.40	47.60	53.80	47.60	40.20	42.80	58.40	54.20	56.80	66.80	65.60	66.80
LogiQA2	30.34	28.75	30.79	36.07	30.79	35.08	40.20	35.18	40.52	38.36	31.87	37.45	42.88	37.47	40.14
GSM8K	40.56	35.18	40.11	83.62	73.92	81.45	76.50	74.83	78.70	78.17	67.55	74.55	79.83	79.30	85.44
DROP	12.32	6.12	9.80	16.01	13.39	16.94	16.22	7.13	11.69	12.05	7.15	11.53	20.60	20.60	26.41
CoQA	57.57	55.93	59.42	78.40	65.23	80.04	78.74	73.63	78.80	77.83	77.50	80.48	78.20	74.44	77.51
SocialIQA	40.38	39.00	40.33	49.59	46.11	50.06	51.59	50.20	55.89	49.85	49.74	49.95	54.04	51.23	55.22
GPQA	29.69	27.90	29.12	30.58	30.13	31.70	34.38	32.37	34.38	34.60	31.03	33.81	36.83	35.04	36.16
ARC-C	34.39	34.04	36.09	58.45	58.36	58.70	55.29	52.65	54.10	53.41	53.24	53.33	60.41	60.41	61.26
Text Avg.	**29.32**	**26.72**	**29.36**	**45.71**	**39.88**	**45.08**	**44.86**	**39.71**	**43.83**	**44.01**	**39.65**	**43.29**	**48.79**	**44.51**	**49.27**
Multimodal Benchmarks															
MMBench-EN	–	52.58	52.52	–	50.52	50.89	–	64.78	64.64	–	60.14	61.94	–	75.17	74.31
RealWorldQA	–	45.10	44.75	–	45.10	44.31	–	52.03	52.01	–	26.67	44.97	–	56.34	55.64
MMMU	–	30.67	32.67	–	36.78	36.22	–	41.00	42.56	–	35.89	36.22	–	47.67	48.11
POPE	–	85.42	84.26	–	77.16	77.82	–	87.12	86.54	–	82.23	80.90	–	87.27	87.67
MMStar	–	37.23	37.07	–	30.58	31.26	–	40.56	40.37	–	36.54	33.94	–	50.26	50.14
MM Avg.	–	**50.20**	**50.25**	–	**48.03**	**48.10**	–	**57.10**	**57.22**	–	**48.29**	**51.59**	–	**63.34**	**63.17**

multimodal adaptation. Consequently, multimodal performance gradually declines at larger values of τ_a. The results support our default choice of $\tau_a = 0.8$, which achieves a favorable balance between language preservation and multimodal adaptation.

Probing Set Size. We further examine the sensitivity of NeuPAT to the number of samples in each probing set. As shown in Table 9, the overall performance is relatively stable across different sizes, with variations of only 0.52 and 0.36 points in the text and multimodal averages, respectively. Increasing N_a from 512 to 2048 generally improves performance, and $N_a = 2048$ achieves the best averages on both language and multimodal benchmarks. Further increasing the size to 4096 provides no additional benefit, indicating diminishing returns from additional probing samples. The results support $N_a = 2048$ as a reasonable default while showing that NeuPAT is relatively insensitive to the probing-set size.

High Response Neuron Regularization Coefficients. We study the sensitivity to the high response neuron regularization coefficients by setting $\lambda_{\text{in}} = \lambda_{\text{out}}$. As shown in Table 10, a weak constraint of 0.05 achieves the highest multimodal average but lower language performance. Increasing the coefficients beyond 0.1 provides no further text improvement and gradually reduces multimodal performance. These results support our default setting of $\lambda_{\text{in}} = \lambda_{\text{out}} = 0.1$, which provides a favorable text-multimodal trade-off.

D.2 Complete Cross-Backbone Results

Due to space limitations, the main text reports results on five representative language benchmarks. Table 11 provides the complete results on all 11 language and 5 multimodal benchmarks. Vanilla Tuning consistently degrades the average language performance across all tested language backbones. In contrast, NeuPAT improves over Vanilla Tuning by 2.64, 5.20, 4.12, 3.64, and 4.76 points on Qwen3-0.6B, Phi-4-Mini-Instruct, Qwen2.5-7B-Instruct, Llama3.1-8B-Instruct, and Qwen2.5-14B-Instruct, respectively. It also recovers the original LLM average within approximately one point for all backbones and slightly surpasses it on Qwen3-0.6B and Qwen2.5-14B-Instruct. Meanwhile, multimodal performance remains compa-

Table 12: Complete ablation results on 11 language and 5 multimodal benchmarks. Variants are grouped by global update strategy, neuron allocation, and neuron-wise plasticity allocation strategies. The complete NeuPAT configuration is repeated and highlighted in each group for comparison.

Variant	Language Benchmarks												Multimodal Benchmarks					
	SIQA	ARC-C	GPQA	MATH	GSM8K	LogiQA2	BBH	bAbI	MBPP	DROP	CoQA	Avg.	MMB	RWQA	MMMU	POPE	MMStar	Avg.
(a) Global vs. Neuron-Aware Update																		
Global Freeze	49.33	56.06	37.72	61.20	86.05	34.29	53.57	13.65	67.10	14.90	70.37	49.48	68.64	53.99	41.33	86.03	43.08	58.61
Uniform Update (Vanilla Tuning)	46.78	55.72	34.15	50.40	83.47	32.95	37.12	2.68	62.60	9.90	67.37	43.92	72.08	56.60	43.11	87.69	45.40	60.98
Global Reg.	49.33	56.14	36.83	54.40	84.99	34.73	55.23	9.33	65.50	9.68	70.15	47.85	72.08	54.25	42.56	86.03	43.91	59.77
NeuPAT	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
(b) Neuron Allocation Strategy																		
Random Partition	49.33	55.55	36.16	55.40	83.69	34.41	54.32	1.94	65.40	9.67	69.42	46.84	69.76	54.51	43.33	87.52	44.47	59.92
Fixed-ratio Partition	50.26	56.23	35.49	53.20	84.90	35.94	52.85	10.24	66.20	10.11	69.85	47.75	72.16	57.25	41.33	87.28	43.26	60.26
NeuPAT	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
(c) Neuron-Wise Update Constraint																		
w/o Text Freeze	49.80	55.29	35.71	52.60	83.90	36.64	52.96	5.87	64.80	10.39	70.23	47.11	72.11	57.30	42.33	87.20	45.68	60.92
w/o Vision Update	50.41	55.63	36.16	53.20	84.44	36.45	53.99	11.85	67.10	12.77	70.20	48.38	69.69	55.16	42.44	86.61	44.04	59.59
w/o Low Response Update	49.74	56.31	37.28	53.00	84.76	35.88	53.89	11.64	66.80	9.96	70.13	48.13	71.05	56.34	42.67	86.94	45.72	60.54
High Response Full Update	48.41	55.46	34.15	53.60	83.61	37.47	52.40	9.94	64.00	11.99	67.82	47.17	72.34	57.25	42.78	87.27	44.72	60.87
High Response Freeze	49.90	56.66	37.05	55.80	85.06	35.37	53.29	15.16	67.20	13.70	70.48	49.06	69.93	55.03	42.67	87.30	45.10	60.01
NeuPAT	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05
(d) High Response Neuron Regularization Design																		
l2-l2	50.00	55.57	37.95	55.20	85.22	36.13	52.97	11.41	66.20	13.06	69.45	48.47	72.20	57.65	42.00	87.84	46.07	61.15
cos-cos	50.05	56.66	37.28	54.20	84.99	36.51	52.34	8.44	67.40	14.51	69.62	48.36	72.16	56.99	42.78	87.37	44.50	60.76
cos-l2	50.00	56.91	36.38	54.60	85.22	36.07	53.26	15.60	67.00	13.40	69.75	48.93	72.51	55.29	43.44	87.14	45.26	60.73
NeuPAT (l2-cos)	50.26	56.91	37.27	56.40	85.97	36.39	53.88	10.67	66.60	15.74	69.53	49.06	72.48	56.99	42.22	88.72	44.85	61.05

rable to Vanilla Tuning, with particularly notable gains on Llama3.1-8B-Instruct. Although the improvements vary across individual tasks, the overall results confirm that NeuPAT generalizes across different model families and scales.

D.3 Complete Ablation Results

Complete Results for the Main-Text Ablations. Table 12 reports the complete results for the three ablation groups presented in the main text. Global update strategies reveal the trade-off between language preservation and multimodal adaptation. Alternative partitioning strategies verify the importance of neuron adaptation role allocation, while the update-strategy ablations demonstrate the role of each neuron set. The complete results are consistent with the conclusions drawn from the representative benchmarks in the main text.

High Response Neuron Regularization Design. We further compare different input- and output-side regularization combinations for high response neurons. As shown in Table 12(d), l2-l2 achieves slightly higher multimodal performance but lower text performance, while cos-cos and cos-l2 yield weaker overall trade-offs. The proposed l2-cos design obtains the highest text average while maintaining near-best multimodal performance, providing the most balanced result. This suggests that input-side parameters benefit from magnitude constraints, whereas output-side parameters are better regularized by preserving their transformation directions.

D.4 Additional Neuron Visualizations

D.4.1 Neuron Distribution across LLM Backbones

Figure 5 compares the layer-wise distributions of the neuron sets identified by importance-guided allocation across different LLM backbones. Although their proportions vary across model families, scales, and layers, text-critical, vision-critical, high response, and low response neurons consistently coexist throughout the backbone, with high response neurons generally forming the largest set. Some models exhibit stronger fluctuations in early layers, while later layers tend to show more stable distributions. These consistent yet heterogeneous patterns suggest that neuron-wise adaptation dynamics generalize across backbones, supporting the architecture-agnostic applicability of NeuPAT.

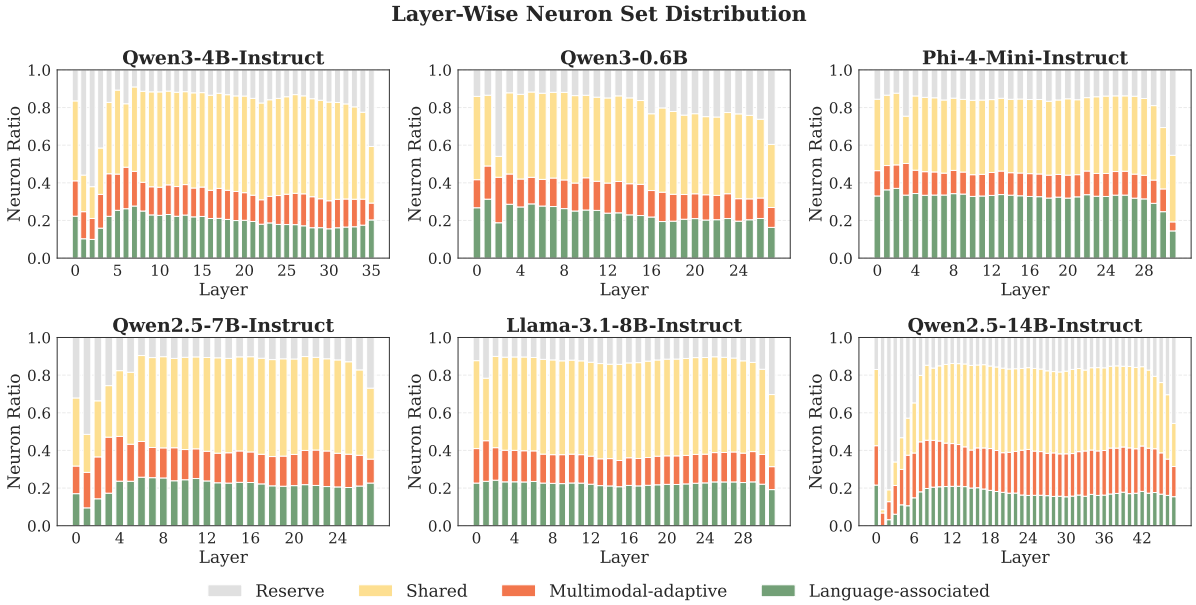

Figure 5: **Layer-wise neuron distributions across LLM backbones.** High response neurons form the largest group in most layers, while the others vary across models and depths.