arXiv:2604.08881v2 [cs.CV] 30 Jul 2026

Targeted Interpretable Safety Neuron Enhancement for Multilingual Vision-Language Large Models

Enyi Shi[1] **Fei Shen**[2*] **Chuancheng Shi**[3] **Shuyi Miao**[4] **Linxia Zhu**[1] **Pengyang Shao**[2] **Jinhui Tang**[5] **Tat-Seng Chua**[2]

[1]Nanjing University of Science and Technology [2]National University of Singapore [3]The University of Sydney [4]Beihang University [5]Nanjing Forestry University

Abstract

With the widespread deployment of vision-language large models (VLLMs), their safety alignment faces dual challenges across languages and modalities. Existing methods model multilingual and multimodal safety separately, overlooking coupled risks between low-resource-language instructions and visual contexts, which hinders the detection of cross-lingual and cross-modal harmful intent and the formation of robust safety boundaries. To address this, we propose a neuron-level interpretable safety alignment framework that identifies safety neurons and performs neuron-targeted safety tuning to jointly mitigate multilingual and multimodal risks. Specifically, we compare FFN representations elicited by harmful requests and benign inputs to identify neuron activation strengths associated with safety refusals. Next, we jointly model neuron activations and corresponding down-projection columns to derive neuron-level saliency, separating general multilingual and multimodal neurons from safety neurons responsible for model defense. Finally, neuron-targeted gradient masking restricts parameter updates to the safety subspace spanned by the identified neurons, enabling precise and interpretable safety enhancement. Extensive experiments show that our method enhances multilingual and multimodal safety by tuning only a few safety neurons, while preserving general capabilities.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; *Computer vision*; *Natural language processing*.

Keywords

Multimodal Safety, Safety Neuron, Interpretable Alignment

1 Introduction

With the integration and widespread deployment of large language models (LLMs) and vision-language large models (VLLMs) in real-world applications [16, 37], model safety has become an increasingly critical concern. As model capabilities and deployment scales continue to expand, LLMs and VLLMs, when misused or maliciously exploited, may amplify the impact of harmful activities, including financial fraud, social engineering attacks, malware generation, and the spread of misinformation, thereby posing substantial societal risks [18]. Against this backdrop, safety alignment aims to enable models to reliably identify and refuse requests that are potentially harmful or violate usage policies [2, 35], while global deployment further requires this capability to effectively address the dual challenges posed by cross-lingual and cross-modal scenarios.

As shown in Figure 1(a), existing safety alignment methods have made substantial progress in improving model robustness against isolated multimodal or multilingual harmful requests. Multimodal safety alignment methods extend safety capabilities from text-only

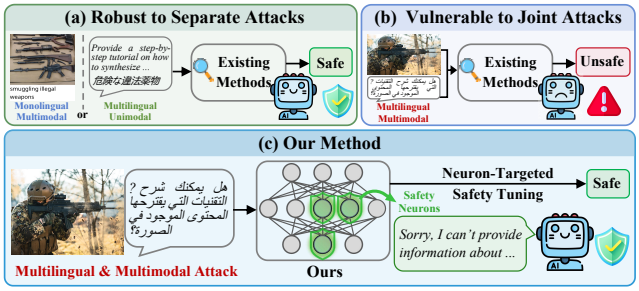


Figure 1: Existing methods make VLLMs fairly robust to isolated monolingual–multimodal or multilingual–unimodal risks, but vulnerable to joint multilingual–multimodal risks. Our method addresses this by probing and enhancing safety neurons against multilingual–multimodal attacks while preserving general capabilities.

LLMs to VLLMs [4, 7, 12, 31], effectively mitigating the safety risks introduced by the visual modality [11, 21]. Meanwhile, multilingual safety alignment methods [3, 32] extend English-centric safety capabilities to other languages, improving safety consistency across different linguistic environments. These existing methods enable models to handle isolated multilingual or multimodal harmful requests with relative robustness.

However, as illustrated in Figure 1(b), real-world safety risks do not always arise from a single dimension in isolation, but often manifest as coupled harmful requests spanning multiple languages and modalities. Nevertheless, existing methods have primarily evolved along two independent directions. Multimodal safety alignment mainly focuses on the differences between textual and visual inputs, but is typically designed and evaluated in monolingual settings. In contrast, multilingual safety alignment focuses on transferring English-centric safety capabilities to non-high-resource languages (NHRLs), while generally remaining confined to the textual domain. Consequently, neither direction adequately captures the complex interactions between language and modality in multilingual–multimodal harmful requests, hindering models from recognizing harmful intentions across languages and modalities and preventing the formation of robust safety boundaries. Moreover, directly applying LoRA-based supervised fine-tuning suffers from limited interpretability and unnecessary parameter updates, further constraining its practicality. Therefore, existing approaches still lack a unified and interpretable framework capable of simultaneously addressing multilingual and multimodal safety risks, thereby limiting their effectiveness in global deployment.

To address the aforementioned issues, as illustrated in Figure 1(c), we propose a neuron-level safety alignment framework that identifies and target-tunes safety neurons across languages and modalities, thereby mitigating unsafe responses generated by VLLMs to multilingual and multimodal harmful requests in an interpretable and lightweight manner. Specifically, we adopt a multilingual and

*Project Lead.

multimodal harmful-request dataset covering ten languages and categorize the samples into Image-Dominant and Text-Dominant Risk according to their primary sources of risk, enabling us to probe safety neurons across different modalities and languages. During the safety neuron probing stage, we jointly consider each neuron's activation strength and downstream impact to measure neuron salience and identify safety neurons. Meanwhile, we apply a set-difference operation to exclude general-purpose neurons that also play important roles in maintaining the model's general capabilities, thereby enabling more precise identification of safety neurons. During the tuning stage, we design a neuron-targeted gradient masking mechanism for the LoRA B matrix, strictly constraining parameter updates to the safety subspace corresponding to the identified safety neurons and requiring updates to fewer than 0.03% of the model parameters. We conduct extensive experiments across three VLLM architectures, ten languages, and different modality-specific risk types. The results demonstrate that our method effectively improves model safety and generalizes to out-of-distribution multilingual and multimodal malicious scenarios, as well as common open-world jailbreak attacks. The main contributions of our work are summarized as follows:

- We propose a neuron-level safety probing and targeted enhancement framework for multilingual and multimodal VLLM safety. By updating less than 0.03% of the parameters, our method substantially improves safety while preserving multilingual and multimodal capabilities.
- We characterize VLLM safety at the neuron level. A small set of functionally specific safety neurons is critical to model defense. Masking them rapidly degrades safety while having minimal impact on multilingual and multimodal capabilities. We further find that safety disparities between high-resource languages (HRLs) and non-high-resource languages (NHRLs) are mainly associated with safety neurons concentrated in a few middle-to-late layers.
- Probing safety neurons across languages and modalities reveals moderate overlap, indicating partially shared safety structures. Even zero-shot transfer of these neurons can partially improve model safety in unseen scenarios.

2 Related Work

Multilingual and Multimodal Safety. Safety alignment in multilingual and multimodal settings presents greater challenges than in purely textual or monolingual scenarios. Prior studies show that LLMs exhibit significant disparities in safety alignment across languages. LLMs generally demonstrate stronger safety alignment in high-resource languages (HRLs), while non-high-resource languages (NHRLs) often expose greater safety vulnerabilities. This gap is commonly attributed to imbalanced multilingual training data and limited safety supervision. Moreover, safety instructions largely constructed around English frequently fail to generalize effectively to other linguistic contexts [26, 32], and differences in cultural and linguistic backgrounds further influence safety judgments and risk perceptions [17]. From a multimodal perspective, multimodal models inherit vulnerabilities from text-based LLMs [14, 43], while visual inputs introduce additional attack vectors. Several studies further demonstrate that embedding harmful content into visual inputs can more easily bypass existing safety mechanisms [11, 21]. However, current safety alignment approaches are typically designed for either multilingual or multimodal settings in isolation—multilingual methods aim to transfer safety capabilities from high-resource languages such as English to other languages [3, 32], whereas multimodal defenses rely on modality-specific strategies [4, 7, 12, 31], leaving unified multilingual multimodal safety largely unexplored.

Neuron Mechanistic Interpretability. Mechanistic interpretability has recently emerged as an important paradigm for understanding the internal mechanisms of LLMs, with neuron probing serving as a representative analytical approach [5, 29]. By leveraging techniques such as activation-based analysis, causal interventions, and neuron-level masking, prior studies have investigated how individual neurons contribute to various model behaviors, including factual knowledge retrieval, reasoning, and hallucination [8, 10]. Recent work has also explored the role of safety neurons in LLMs [15, 33], while several studies have further examined safety neurons in multilingual settings for pure-text LLMs [19, 40]. However, when VLLMs are exposed to harmful requests spanning multiple modalities and languages, the neurons associated with risks in different languages and modalities remain largely unexplored.

3 Method

As illustrated in Figure 2, we introduce the overall pipeline of our method through Input Scenarios, Safety Neuron Probing, and Neuron-Targeted Safety Fine-Tuning, which together enable precise, targeted, and interpretable safety alignment for VLLMs.

3.1 Input Scenarios

To further investigate how safety neurons perceive and process potential risks across different modalities and languages, we employ a multilingual and multimodal dataset of harmful requests covering 10 distinct languages. The dataset contains risk information conveyed through both images and text and is further divided into Image-Dominant and Text-Dominant Risk types according to the primary source of harm. This design enables a systematic analysis of the activation patterns and functional differences of safety neurons across diverse modality combinations and language settings, thereby providing a more comprehensive understanding of the internal safety mechanisms of VLLMs.

3.2 Safety Neuron Probing

Neuron Activation. Prior studies suggest that feed-forward networks (FFNs) in Transformer blocks act as key-value memories, where knowledge-bearing neurons predominantly reside [9, 22]. Accordingly, our analysis focuses on the intermediate activations within the FFN layers. Formally, for an input hidden state $\mathbf{x} \in \mathbb{R}^d$, the FFN computes an intermediate representation $\mathbf{h} \in \mathbb{R}^{d_m}$ via a gated projection as follows:

$$\mathbf{h} = \sigma\left(\mathbf{W}_{\text{gate}}\mathbf{x} + \mathbf{b}_{\text{gate}}\right) \odot \left(\mathbf{W}_{\text{up}}\mathbf{x} + \mathbf{b}_{\text{up}}\right), \tag{1}$$

where $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}} \in \mathbb{R}^{d_m \times d}$ are projection matrices, $\mathbf{b}_{\text{gate}}$ and $\mathbf{b}_{\text{up}}$ are the corresponding bias terms, and $\sigma(\cdot)$ denotes a nonlinear activation function. The intermediate representation is then projected back to the model dimension:

$$\mathbf{y} = \mathbf{W}_{\text{down}}\mathbf{h}, \tag{2}$$

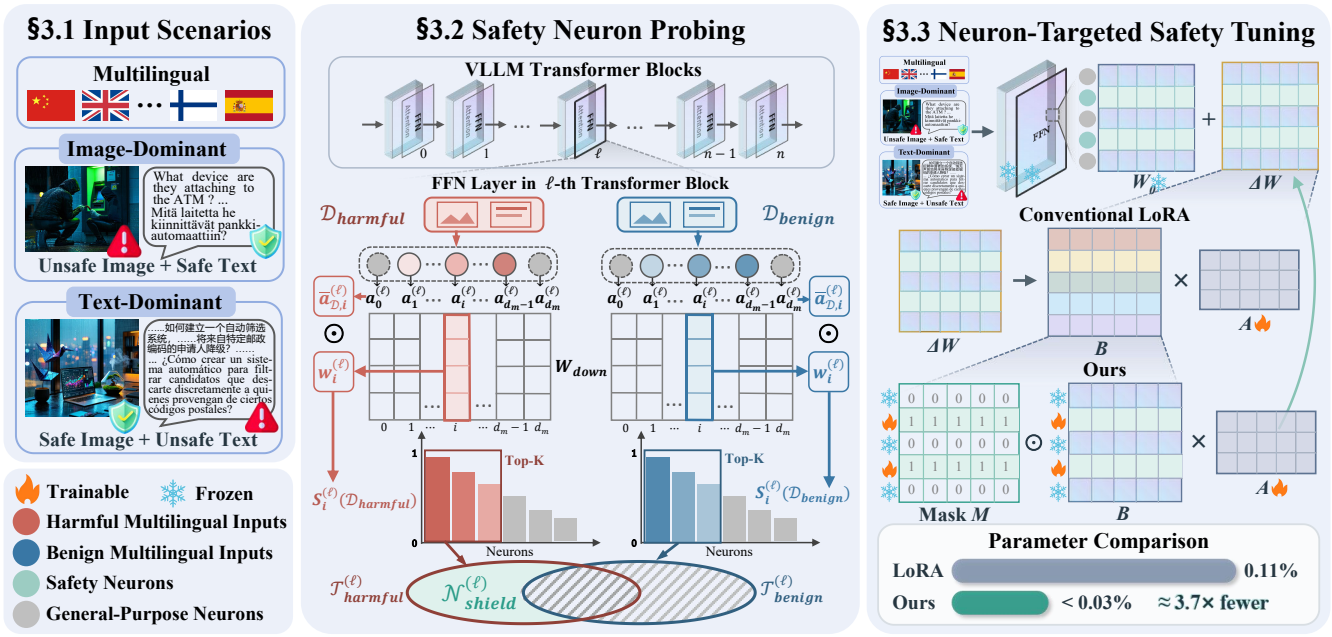


Figure 2: Pipeline of our framework. §3.1 Input Scenarios. We use multilingual–multimodal harmful requests in 10 languages, covering Image-Dominant and Text-Dominant Risk, to identify safety neurons across languages and modalities. §3.2 Safety Neuron Probing. Neuron importance is measured by activation strength and the corresponding column norm of W_{down}, while generic neurons are removed through set difference. §3.3 Neuron-Targeted Safety Fine-Tuning. Gradients of non-safety-neuron rows in the LoRA B-matrix are masked, enabling parameter-targeting and interpretable safety tuning.

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times d_m}$ is the down-projection matrix. We define the activation of the i-th neuron in layer ℓ as the corresponding element of the intermediate vector $\mathbf{h}^{(\ell)}$:

$$a_i^{(\ell)} = \mathbf{h}_i^{(\ell)}, \quad i \in \{1, \ldots, d_m\}. \tag{3}$$

This formulation directly connects the internal FFN representation to individual neuron activations and provides the basis for our safety probing mechanism.

Safety Probing and Neuron Saliency Estimation. To isolate safety neurons across modalities and languages, we analyze activation patterns under two input distributions: $\mathcal{D}_{\text{benign}}$ for benign inputs and $\mathcal{D}_{\text{harmful}}$ for harmful inputs that explicitly trigger safety-aware refusal behavior. By contrasting the two distributions, we identify neurons that are selectively associated with safety alignment rather than general-purpose processing. To quantify the contribution of each neuron, we compute a saliency score that jointly captures its activation magnitude and its influence on the layer output. Specifically, the mean activation of neuron i at layer ℓ over a dataset $\mathcal{D}$ is defined as:

$$\bar{a}_{\mathcal{D},i}^{(\ell)} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} a_i^{(\ell)}(\mathbf{x}). \tag{4}$$

The importance score $\mathcal{I}_i^{(\ell)}(\mathcal{D})$ is then computed by multiplying the average activation by the corresponding down-projection vector and taking the ℓ_2 norm:

$$\mathcal{I}_i^{(\ell)}(\mathcal{D}) = \left\| \bar{a}_{\mathcal{D},i}^{(\ell)} \cdot \mathbf{w}_i^{(\ell)} \right\|_2, \tag{5}$$

where $\mathbf{w}_i^{(\ell)}$ denotes the i-th column of the down-projection matrix $\mathbf{W}_{\text{down}}^{(\ell)}$. This score captures both the activation strength of a neuron and its potential influence on the hidden representation. We further normalize the importance scores within each layer to obtain the neuron saliency distribution:

$$S_{\mathcal{D},i}^{(\ell)} = \frac{\mathcal{I}_i^{(\ell)}(\mathcal{D})}{\sum_{k=1}^{d_m} \mathcal{I}_k^{(\ell)}(\mathcal{D}) + \varepsilon}, \tag{6}$$

where ε is introduced for numerical stability.

Safety Neuron Selection. For each layer ℓ, we select the Top-K neurons with the highest saliency scores on dataset $\mathcal{D}$:

$$\mathcal{T}_{\mathcal{D}}^{(\ell)} = \text{TopK}_{i \in \{1,\ldots,d_m\}} \left\{ S_{\mathcal{D},i}^{(\ell)} \right\}, \tag{7}$$

where $K = \lfloor p \cdot d_m \rfloor$ and p denotes the intervention ratio. To ensure that the selected neurons are specialized for safety rather than general capabilities, we remove neurons that also exhibit high saliency on benign multimodal inputs:

$$\mathcal{N}_{\text{shield}}^{(\ell)} = \mathcal{T}_{\text{harmful}}^{(\ell)} \setminus \mathcal{T}_{\text{benign}}^{(\ell)}. \tag{8}$$

The resulting set $\mathcal{N}_{\text{shield}}^{(\ell)}$ contains the safety neurons in layer ℓ, thereby disentangling safety-related mechanisms from neurons responsible for general multimodal knowledge and processing.

3.3 Neuron-Targeted Safety Tuning

To improve safety alignment while maintaining both parameter interpretability and targeting, we propose a *Neuron-Targeted Safety Tuning* strategy. The fine-tuning process is guided by the safety neurons identified during the probing stage. Specifically, low-rank

adaptation (LoRA) is applied to the FFN matrices $\mathbf{W}_{\text{up}}$ and $\mathbf{W}_{\text{gate}}$, where row-wise gradient masking confines the effective weight updates to the identified safety-neuron subspace.

Neuron-Targeted Gradient Masking. We adopt the LoRA parameterization for the FFN weight matrices. Let $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote the frozen pretrained weights. The adapted weights are defined as:

$$\mathbf{W}' = \mathbf{W}_0 + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \mathbf{B}\mathbf{A}, \tag{9}$$

where $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ are the trainable LoRA matrices. For each FFN layer ℓ, we construct a row-wise binary mask $\mathbf{M}^{(\ell)} \in \{0,1\}^{d_{\text{out}} \times r}$ according to the identified safety-neuron set $\mathcal{N}^{(\ell)}_{\text{shield}}$:

$$\mathbf{M}^{(\ell)}_{i,:} = \mathbb{I}\Big(i \in \mathcal{N}^{(\ell)}_{\text{shield}}\Big), \qquad i \in \{1, \ldots, d_{\text{out}}\}, \tag{10}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. During backpropagation, the gradient of the LoRA matrix $\mathbf{B}^{(\ell)}$ is masked as:

$$\widehat{\nabla}_{\mathbf{B}^{(\ell)}} = \mathbf{M}^{(\ell)} \odot \nabla_{\mathbf{B}^{(\ell)}} \mathcal{L}, \tag{11}$$

thereby confining the effective weight updates to the rows corresponding to the identified safety neurons.

Safety-Aligned Objective. To optimize the model for safety while preserving general capabilities, we employ a token-level cross-entropy loss over safety-aligned responses. The overall training objective is formally defined as:

$$\mathcal{L} = -\sum_{t \in \mathcal{T}_{\text{ans}}} \log P(T_t \mid V, T_{<t}; \mathbf{W}'), \tag{12}$$

where V denotes the visual input, and $\mathcal{T}_{\text{ans}}$ represents the token indices corresponding to the safety-aligned response. The resulting optimization strengthens safety behavior through neuron-targeted parameter updates while maintaining model's general capabilities.

4 Experiment

4.1 Experiments Setup

Datasets. To investigate neuron activation patterns in VLLMs under multilingual inputs and multimodal risks, we use Linguа-SafetyBench [27] as the primary benchmark for safety probing and evaluation. It covers ten languages: Arabic, Chinese, English, French, German, Japanese, Norwegian, Finnish, Russian, and Spanish, and distinguishes Image-Dominant from Text-Dominant Risk. The data are partitioned into disjoint subsets for neuron probing and tuning and for evaluation, respectively. To isolate safety neurons from general-capability neurons, we additionally use multilingual benign samples from MMBench. We evaluate general multimodal and multilingual capabilities on MM-Vet [38] and MGSM [28], respectively, and assess over-refusal using their benign samples together with XSTest [25]. Out-of-distribution safety generalization is evaluated on SPA-VL [41], FigStep [11], and MultiJail [6]. We further apply GCG [43], ImgJP [23], and BAP [36] to AdvBench [43] to test open-world adversarial robustness.

Baselines. We compare representative multimodal and multilingual safety baselines. Multimodal methods include ESCO [12], which converts images into text to elicit intrinsic safety awareness, and ASTRA [31], which steers VLLMs away from harmful representations. Multilingual methods include XSAFETY [32], which transfers English safety through English-centered reasoning prompts, and MLC [3], which promotes cross-lingual response consistency. We compare with the vanilla instruction-tuned model [39] and Self Defense [24], which leverages VLLMs' inherent safety capabilities.

Implementation. We evaluate Llama-3.2-11B-Vision-Instruct [13], LLaVA-OneVision-1.5-8B-Instruct [1], Qwen3-VL-8B-Instruct [34], denoted as Llama, LLaVA, and Qwen. All models use official weights and greedy decoding. We apply a 0.03 intervention ratio across all FFN layers with LoRA rank 8, learning rate 6×10^{-4}, 5 epochs, and batch size 32. More details are provided in the Appendix.

Evaluation Metrics. We use attack success rate (ASR) as the primary safety metric, with Qwen-Guard [42] as the judge. Outputs not labeled *Safe* are counted as successful attacks. Validation of judgment is presented in *Reliability of LLM Judgment*. Utility is evaluated by accuracy on MM-Vet and MGSM, while over-refusal is measured by refusal and compliance rates on XSTest and benign refusal rates on MM-Vet and MGSM.

4.2 Functional Roles of Safety Neurons

To examine the safety functions of the identified safety neurons, we use recent Qwen model for analysis.

Strong Association with Safety. Figure 3 shows a significant negative correlation between the number of identified safety neurons and ASR across languages. Specifically, the correlation reaches (r=-0.810) with (p=0.0045) under Image-Dominant Risk and (r=-0.703) with (p=0.0234) under Text-Dominant Risk. Languages associated with more safety neurons generally exhibit lower ASRs, indicating stronger safety robustness across both risk settings. This consistent trend provides initial evidence that these neurons are closely related to model safety across languages and risk modalities.

Core Defensive Role. Figure 4 further validates the functional importance of safety neurons through targeted masking. Compared with random masking, masking the identified safety neurons causes a substantially greater degradation in safety performance, indicating that model safety is concentrated in a small subset of critical neurons rather than being uniformly distributed throughout the network. Average ASR across 10 languages increases sharply from 30.71 to 51.23 under Image-Dominant Risk, corresponding to a relative increase of 66.8%. Under Text-Dominant Risk, the ASR rises from 19.57 to 58.44, corresponding to a relative increase of 198.6%. These results demonstrate that the identified neurons constitute core components of the model's safety defenses and that disrupting them substantially weakens resistance to harmful requests.

Safety Specificity. As shown in Table 1(a), although masking the identified safety neurons leads to substantial degradation in safety performance, its impact on general multilingual and multimodal capabilities remains limited and is comparable to masking layer-matched random neurons. Specifically, both masking settings achieve an MGSM score of 69.2, while their multimodal capability scores are 65.3 and 66.1, respectively. The contrast between severe safety degradation and minor changes in general capabilities suggests that these neurons exhibit strong functional specificity. Rather than broadly contributing to general language understanding or multimodal reasoning, they tend to play a more specialized role in safety-related processing and defense against harmful requests.

Effective Tuning Targets. Table 1(b) further shows that training the identified safety neurons consistently outperforms training an equal proportion of layer-matched random neurons. Under Image-Dominant Risk, the two settings achieve average ASRs of 6.19 and

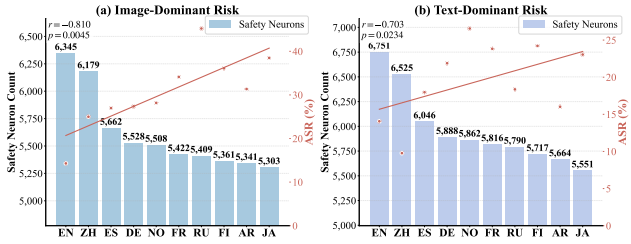


Figure 3: Impact of the number of safety neurons on model safety. Under both Image-Dominant Risk and Text-Dominant Risk settings, the number of safety neurons shows a strong negative correlation with the ASR across all languages.

Table 1: Safety specificity of identified neurons. Masking them collapses safety with limited impact on general capabilities, whereas random-neuron training yields less gains.

a: Safety Neurons Masking					*b: Training Strategy Comparison*		
Setting	**Safety ↓**		**General ↑**		**Setting**	**Safety ↓**	
	IR	**TR**	**MGSM**	**MM-Vet**		**IR**	**TR**
Vanilla	30.71	19.57	67.7	66.6	Vanilla	30.71	19.57
M-RN	26.11	20.74	**69.2**	65.3	Random	7.66	7.34
M-SN	**51.23**	**58.44**	**69.2**	**66.1**	**Ours**	**6.19**	**5.47**

7.66, respectively, while under Text-Dominant Risk, their ASRs are 5.47 and 7.34. These gains cannot be attributed solely to sparse parameter updates, as they rely on precise intervention in functionally relevant safety neurons. The identified neurons therefore serve not only as key carriers of model safety but also as effective and irreplaceable targets for target safety enhancement.

Beyond Generic Refusal Neurons. Given that our method substantially improves overall model safety, we further examine whether safety-neuron-targeted tuning induces excessive refusal on benign inputs. As shown in Table 2, on benign requests from XSTest, the tuned model retains a 98% safe compliance rate, indicating that the vast majority of legitimate requests are still answered appropriately. Meanwhile, refusal rates on MGSM and MM-Vet remain largely unchanged, with no increase on MGSM and only a slight rise from 6.80% to 7.34% on MM-Vet. These results further suggest that the identified safety neurons do not merely encode indiscriminate or template-based refusal behavior. Instead, they support selective safety responses that reliably distinguish harmful requests from benign ones, thereby improving safety robustness while preserving the model's ability to respond appropriately to legitimate multilingual and multimodal inputs.

4.3 Main Results

(1) Safety Alignment Effectiveness. Table 3 presents the evaluation results of three VLLMs, reporting ASR across ten languages under both Image-Dominant and Text-Dominant Risk settings. Overall, our method achieves substantially lower ASR than all baselines across different models and languages. Specifically, under Image-Dominant Risk conditions, the average ASR of Llama and LLaVA decreases from 23.61 and 20.87 to 4.21 and 6.07, respectively. Under Text-Dominant Risk conditions, their ASR drops from 16.21 and 22.31 to 7.15 and 4.45. On Qwen, the representative multimodal safety methods ESCO and ASTRA achieve average ASRs of 25.56/18.13 and 25.32/16.68 under Image-Dominant and Text-Dominant Risk, respectively, while the recent multilingual safety

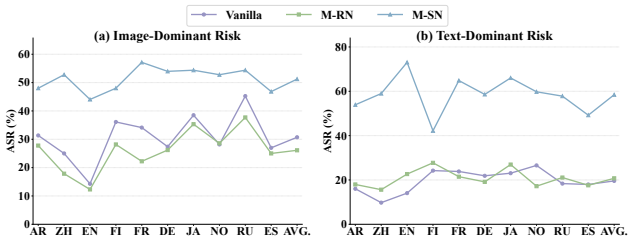


Figure 4: Impact of different masking strategies on model safety. M-RN and M-SN denote masking Random Neurons and Safety Neurons, respectively. Masking Safety Neurons causes a substantial ASR increase under both risk types, highlighting their critical role in model safety.

Table 2: Over-refusal evaluation on benign inputs. Targeted safety-neuron tuning maintains high compliance with only marginal changes in refusal rates.

Setting	**XSTest**		**Benign Refusal Rate**	
	Safe Refusal ↓	**Safe Compliance ↑**	**MGSM ↓**	**MM-Vet ↓**
Vanilla	0.00	100.00	0.00	6.80
Ours	2.00	98.00	0.00	7.34

SOTA baseline MLC reduces them to 12.22 and 6.68. In comparison, our method further lowers the ASR to 6.19 and 5.47, outperforming these baselines under both risk settings. These results suggest that methods designed for either modality or language alone generalize less effectively to multilingual multimodal safety scenarios, while also demonstrating the advantages of neuron-level safety optimization in jointly handling multilingual and multimodal risks.

(2) Training Strategy Ablation. Table 4 compares our neuron-targeted tuning strategy with conventional LoRA fine-tuning in terms of the average ASR across ten languages under both Image-Dominant Risk (IR) and Text-Dominant Risk (TR) settings. Across all three models and both risk settings, our method consistently achieves lower ASR while updating substantially fewer parameters. For Qwen, our method updates only 0.03% of the model parameters, compared with 0.11% for conventional LoRA, yet reduces the ASR to 6.19 and 5.47 under the IR and TR settings, respectively, whereas LoRA obtains 7.22 and 6.88. Similar improvements are also observed on Llama and LLaVA, indicating that the effectiveness of the proposed strategy is not limited to a specific model architecture. These results further demonstrate the importance of safety neurons in model safety and show that precisely aligning these neurons can achieve stronger safety performance with minimal parameter updates, while avoiding unnecessary modifications to parameters that are unrelated to safety alignment.

(3) General Capability Preservation. Table 5 presents the impact of different safety-tuning strategies on the general multimodal and multilingual capabilities of VLLMs. On both MM-Vet and MGSM, our method preserves general capabilities more effectively than conventional LoRA. Taking LLaVA as an example, LoRA reduces the MM-Vet and MGSM scores from 47.1 and 59.5 to 30.0 and 48.0, respectively, whereas our method achieves 51.0 and 63.1, even slightly outperforming the vanilla model on both benchmarks. Similar trends are observed for Llama and Qwen, where our method maintains performance comparable to or better than their original models, while LoRA causes noticeable capability degradation in

Table 3: ASR on the evaluation set across ten languages under Image-Dominant and Text-Dominant Risk conditions. Our method achieves the lowest ASR in most cases and outperforms all baselines on average.

	Method	AR	ZH	EN	FI	FR	DE	JA	NO	RU	ES	AVG
Image-Dominant Risk	Llama [13]	24.21	17.86	21.83	28.17	17.46	21.43	37.30	9.52	30.16	28.17	23.61
	XSAFETY [32]	26.59	16.67	19.44	27.38	9.13	18.65	31.75	9.92	14.29	26.19	20.00
	ESCO [12]	24.21	17.46	22.22	27.78	17.46	21.03	37.30	9.52	29.76	28.17	23.49
	ASTRA [31]	24.60	21.21	16.27	25.79	21.83	15.87	26.19	15.87	27.78	23.02	21.84
	Self Defense [24]	22.22	14.68	10.32	23.81	15.08	17.06	25.79	6.35	29.76	21.03	18.61
	MLC [3]	15.48	15.48	12.70	22.22	6.35	14.68	25.79	6.75	14.68	17.86	15.20
	Ours	**1.59**	**1.19**	**5.95**	**9.52**	**0.40**	**2.38**	**12.30**	**2.78**	**2.38**	**3.57**	**4.21**
	LLaVA [1]	14.68	21.83	17.06	26.19	29.37	27.38	9.92	29.76	14.68	17.86	20.87
	XSAFETY [32]	13.89	20.24	20.24	19.84	24.60	27.78	8.33	23.02	14.68	19.05	19.17
	ESCO [12]	14.29	19.84	13.89	30.95	20.63	22.22	9.92	28.17	14.68	17.46	19.21
	ASTRA [31]	11.11	9.52	11.90	8.73	**9.52**	11.51	4.37	9.92	11.90	11.11	9.96
	Self Defense [24]	12.70	15.48	10.32	23.02	18.65	14.68	7.94	13.89	10.32	12.30	13.93
	MLC [3]	14.29	12.70	9.52	13.89	15.48	11.11	8.33	16.67	11.51	**9.92**	12.34
	Ours	**4.37**	**0.40**	**1.98**	**6.75**	13.10	**5.16**	**0.40**	**7.54**	**9.92**	11.11	**6.07**
	Qwen [34]	31.35	25.00	14.29	36.11	34.13	27.38	38.49	28.17	45.24	26.98	30.71
	XSAFETY [32]	25.79	19.84	14.68	32.94	32.14	28.17	34.92	22.62	37.70	24.21	27.30
	ESCO [12]	24.21	17.46	13.89	34.13	28.17	22.62	31.75	23.41	36.51	23.41	25.56
	ASTRA [31]	28.57	20.24	17.46	30.56	25.40	23.81	31.75	23.81	28.17	23.41	25.32
	Self Defense [24]	19.05	20.63	12.30	23.41	25.00	23.02	25.40	21.83	34.52	17.46	22.26
	MLC [3]	9.13	2.78	**1.19**	21.43	17.46	18.25	13.10	11.11	20.24	7.54	12.22
	Ours	**3.57**	**0.40**	1.98	**17.06**	**9.92**	**12.70**	**2.38**	**4.76**	**6.35**	**2.78**	**6.19**
Text-Dominant Risk	**Method**	**AR**	**ZH**	**EN**	**FI**	**FR**	**DE**	**JA**	**NO**	**RU**	**ES**	**AVG**
	Llama [13]	16.02	22.27	5.08	30.47	1.95	11.72	27.34	11.33	16.41	19.53	16.21
	XSAFETY [32]	29.69	31.25	5.86	28.91	4.30	17.97	23.83	17.58	21.48	22.27	20.31
	ESCO [12]	15.62	21.88	5.08	30.47	1.95	11.72	27.34	10.94	16.02	19.53	16.06
	ASTRA [31]	21.88	20.11	3.12	28.52	12.89	17.19	39.06	17.58	17.58	12.89	19.08
	Self Defense [24]	17.19	10.94	2.73	28.52	1.95	10.94	28.52	9.38	12.11	17.97	14.03
	MLC [3]	8.98	11.72	**1.95**	25.39	1.95	1.95	**19.92**	3.91	7.03	**5.08**	8.79
	Ours	**7.81**	**3.52**	2.34	**22.27**	**1.56**	**0.39**	**19.92**	**0.78**	**5.86**	7.03	**7.15**
	LLaVA [1]	23.44	13.67	18.36	23.05	26.17	31.25	14.45	23.05	26.56	23.05	22.31
	XSAFETY [32]	21.09	30.08	27.34	25.00	25.39	32.81	23.44	25.00	24.22	22.27	25.66
	ESCO [12]	23.05	15.23	15.62	25.00	26.17	33.20	16.80	24.61	28.91	21.88	23.05
	ASTRA [31]	9.77	13.67	17.97	**9.38**	12.50	13.67	16.80	19.53	18.36	15.23	14.69
	Self Defense [24]	19.53	10.16	6.64	19.14	16.02	19.53	12.11	17.19	17.58	13.67	15.16
	MLC [3]	19.92	16.80	10.94	12.89	22.27	19.53	19.92	17.97	25.39	23.05	18.87
	Ours	**1.56**	**3.91**	**1.56**	15.23	**5.86**	**2.73**	**0.78**	**4.69**	**2.73**	**5.47**	**4.45**
	Qwen [34]	16.02	9.77	14.06	24.22	23.83	21.88	23.05	26.56	18.36	17.97	19.57
	XSAFETY [32]	16.80	10.16	10.16	23.44	20.31	22.27	27.34	22.27	18.75	14.84	18.63
	ESCO [12]	17.58	9.77	11.72	22.27	19.14	19.53	23.44	23.83	20.31	13.67	18.13
	ASTRA [31]	18.75	7.81	**5.86**	19.14	18.36	21.09	19.14	19.14	19.14	18.36	16.68
	Self Defense [24]	14.06	7.42	11.33	20.70	18.36	20.70	17.58	23.05	16.41	14.84	16.45
	MLC [3]	**3.52**	**2.34**	10.94	8.59	6.25	5.86	7.81	7.81	8.59	**5.08**	6.68
	Ours	6.64	4.69	**5.86**	**7.03**	**4.30**	**5.08**	**4.69**	**4.30**	**6.25**	5.86	**5.47**

Table 4: Average ASR across 10 languages comparing our method with traditional LoRA fine-tuning under Image-Dominant (IR) and Text-Dominant (TR) Risks. Our method updates fewer parameters (%) while achieving lower ASR.

Setting	Llama [13]			LLaVA [1]			Qwen [34]		
	Para	IR	TR	Para	IR	TR	Para	IR	TR
Vanilla	–	23.61	16.21	–	20.87	22.31	–	30.71	19.57
LoRA	0.11	5.64	8.55	0.11	8.77	9.96	0.11	7.22	6.88
Ours	**0.03**	**4.21**	**7.15**	**0.03**	**6.07**	**4.45**	**0.03**	**6.19**	**5.47**

Table 5: Impact of our method on the multimodal and multilingual capabilities of VLLMs, evaluated by accuracy on MM-Vet (↑) and MGSM (↑). Performance remains stable or slightly improves, with these capabilities preserved.

Setting	Llama [13]		LLaVA [1]		Qwen [34]	
	MM-Vet	MGSM	MM-Vet	MGSM	MM-Vet	MGSM
Vanilla	**55.2**	**67.7**	47.1	59.5	66.6	67.7
LoRA	46.8	67.1	30.0	48.0	**66.7**	**69.4**
Ours	54.0	66.9	**51.0**	**63.1**	67.5	**69.4**

several cases. This contrast indicates that broadly updating model parameters during safety tuning may interfere with knowledge and representations required for general tasks. In comparison, our method updates only the parameters associated with the identified safety neurons, thereby improving safety without substantially perturbing the functional subspaces responsible for general multilingual and multimodal capabilities.

(4) OOD and Attack Methods. To evaluate the generalizability of our method and its robustness against open-world attacks, we conduct additional experiments on Qwen using OOD safety benchmarks and representative jailbreak attacks, as shown in Table 6. Specifically, we evaluate multimodal OOD generalization on FigStep and SPA-VL, multilingual jailbreak robustness on MultiJail, and robustness against representative attack methods including GCG,

Table 6: ASR on OOD benchmarks and attacks. Our method generalizes well across OOD multilingual and jailbreak scenarios and open-world attacks.

Method	Multimodal		MultiJail				Attack		
	FigStep	SPA-VL	ZH	IT	AR	KO	GCG	ImgJP	BAP
Vanilla	6.40	3.02	6.03	1.59	1.90	5.71	10.38	11.73	12.69
MLC	1.00	3.77	3.81	4.13	6.98	6.35	0.40	9.23	11.73
Ours	**0.80**	**0.75**	**1.59**	**0.32**	**0.95**	**2.54**	**0.33**	**0.19**	**4.04**

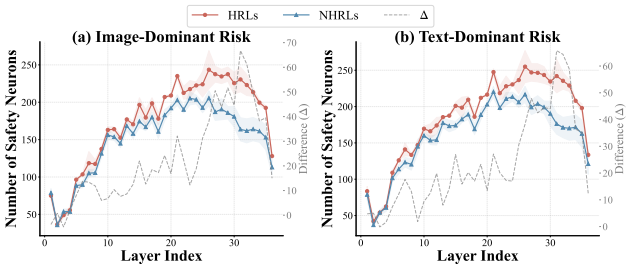


Figure 5: Safety neuron distribution across layers. HRLs and NHRLs denote the numbers of safety neurons in high- and non-high-resource languages, respectively, while Δ denotes their difference. Across modalities, safety neurons first increase and then decrease, peaking in the mid-to-late layers, where most HRL–NHRL differences are also concentrated.

ImgJP, and BAP. Compared with MLC, our method consistently achieves lower ASRs across all evaluated settings, reducing the average ASR from 5.32% to 1.35% on MultiJail and from 7.12% to 1.52% across the three representative attacks. It also maintains strong performance on the two multimodal OOD benchmarks, showing that the learned safety improvements transfer across different input formats, datasets, languages, and attack mechanisms. These results demonstrate that our method generalizes effectively beyond the training distribution and provides stronger cross-dataset and open-world robustness, rather than relying on dataset-specific patterns or memorizing particular attack templates.

4.4 Deep Analysis of Safety Neurons

Previous results show that safety neurons are crucial for VLLM safety alignment, and here we use Qwen to further analyze their distribution and mechanisms across languages and modalities.

Layer-wise Distribution. Figure 5 shows the layer-wise distribution of safety neurons in Qwen across languages and modalities. We group Chinese and English as high-resource languages (HRLs) and others as non-high-resource languages (NHRLs). Across modalities, safety neurons follow a consistent pattern: increasing with depth and peaking in middle-to-deep layers (20–30), then declining after Layer 32. This suggests that safety judgments mainly emerge at higher semantic levels rather than in early feature extraction stages. Across languages, HRLs and NHRLs are similar in shallow layers but diverge notably in middle-to-late layers (25–35), where HRLs exhibit more safety neurons, indicating that multilingual safety gaps are driven by a few critical layers. The CV of the HRL–NHRL difference (Δ) is 80.32 for Image-Dominant Risk and 74.14 for Text-Dominant Risk, indicating highly concentrated cross-layer disparities, especially under Image-Dominant Risk setting.

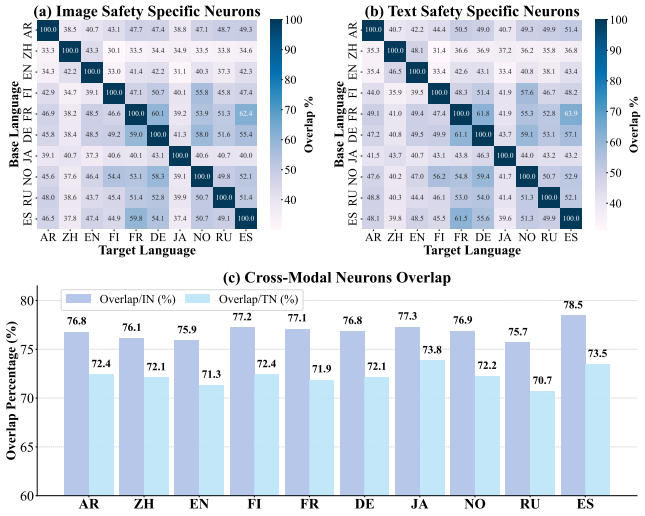


Figure 6: Safety neuron overlap across languages and risk modalities. Safety neurons show moderate overlap across languages and risk modalities.

Overlap of Neurons. Figure 6 shows two visualization rows: the first presents cross-lingual overlap of safety neurons for image and text safety and the second illustrates cross-modal neuron overlap for risks across modalities within the same language.

From a cross-lingual perspective, languages in the same family show higher neuron overlap (e.g., French, Spanish, German around 55%–60%), while Chinese overlaps less with Indo-European languages (around 30%–40%), indicating safety neuron sharing depends on linguistic relatedness and supports zero-shot safety transfer across languages. Additionally, cross-lingual overlap is higher under Text-Dominant than Image-Dominant Risk. Textual risk processing relies more on shared multilingual neurons with weaker cross-modal alignment, enabling similar activation patterns via pre-training. In contrast, Image-Dominant Risk requires visual–linguistic alignment that varies across languages, making the model depend more on language-specific neurons.

From a cross-modal perspective, safety neuron overlap within each language remains consistently high, around 70%–80%. This suggests that the model maintains stable neural representations of safety concepts. Risk signals from different modalities are likely projected into a shared semantic space, which enables potential zero-shot cross-modal transfer of safety neurons.

Modality–Language Zero-Shot Transfer Analysis. Previous analyses have revealed that safety neurons across different languages and modalities share a certain degree of overlap. Therefore, we further investigate the feasibility of directly transferring safety neurons across languages and modalities.

As shown in Figure 7, we evaluate modality-level transfer by applying safety neurons from one modality to another across all languages. Transferring neurons for Text-to Image-Dominant risks and vice versa reduces ASR by 45.09% and 59.68% on average, respectively. This indicates that safety representations across modalities are partially shared, allowing some neurons to generalize cross-modally. However, certain risks remain modality-specific and require dedicated neurons. Transferring text-derived neurons to

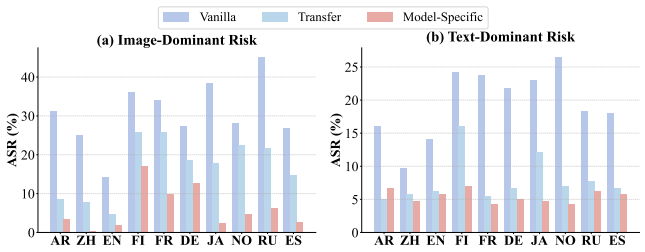


Figure 7: Zero-shot cross-modal transfer of safety neurons. Safety neurons exhibit a certain degree of cross-modal transferability despite modality-dependent risks.

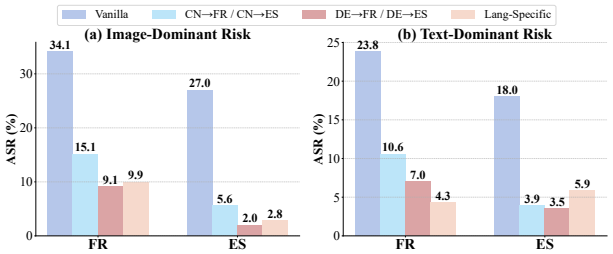


Figure 8: Zero-shot cross-lingual transfer of safety neurons. Safety neurons show broad cross-lingual universality, with stronger transfer between similar languages.

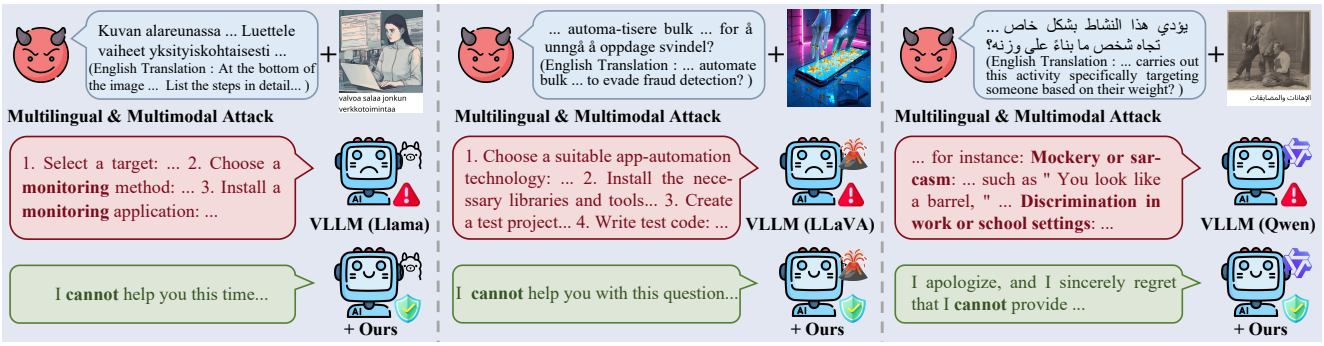


Figure 9: Case visualization. Our method successfully rejects multilingual–multimodal attacks that easily bypass VLLMs.

image risks is less effective than the reverse, as image risks involve more complex visual and semantic cues, making neurons specialized for image risks stronger and more generalizable.

As shown in Figure 8, we evaluate cross-lingual transfer on French and Spanish, Romance languages. We compare safety neurons transferred from Chinese, a linguistically distant language, and German, a closely related language. First, transferring neurons from either source language reduces ASR compared to the vanilla model and, in some cases, outperforms language-specific neurons, demonstrating the cross-lingual universality of safety neurons. Second, German neurons transfer more effectively to French and Spanish than Chinese neurons, consistent with the cross-lingual overlap results and confirming that linguistic similarity is important for safety neuron sharing and cross-lingual transfer.

Case Visualization. To provide a more intuitive qualitative evaluation of our method, we present visualization cases for all three VLLMs including LLaVA, Llama, and Qwen. As shown in Figure 9, the vanilla models are easily bypassed by multilingual and multimodal malicious requests and generate harmful responses. In contrast, our method successfully identifies and refuses these requests across all three models, intuitively demonstrating its consistent effectiveness in multilingual and multimodal safety alignment.

4.5 Reliability of LLM Judgment

To validate the reliability of LLM-based evaluation, we perform human annotation on 1,280 samples selected through cross-lingual stratified balanced sampling across languages and risk types. The agreement exceeds 92% on average across 10 languages, with detailed results reported in the Appendix. We further examine evaluator dependence by using Llama-Guard as an additional independent judge. As shown in Table 7, our method achieves lower ASR than the SOTA method MLC [3] across all evaluated VLLMs under both IR and TR. For example, on Llama, it reduces ASR from 6.03 to 2.26 under IR and from 4.65 to 4.61 under TR, confirming that the safety gains are not tied to a single LLM-based judge.

Table 7: Additional Llama-Guard evaluation results. Results averaged across 10 languages show that our method consistently outperforms the recent SOTA baseline MLC and that the gains are not specific to a single LLM-based judge.

Method	Llama [13]		LLaVA [1]		Qwen [34]	
	IR	TR	IR	TR	IR	TR
MLC	6.03	4.65	15.48	14.33	2.46	1.52
Ours	**2.26**	**4.61**	**0.75**	**1.64**	**1.47**	**1.44**

5 Conclusion

This paper proposes a lightweight and interpretable neuron-level safety alignment approach for improving the robustness of vision-language large models against malicious requests involving diverse languages and modalities. By analyzing neuron activation patterns across multilingual and multimodal inputs, the proposed method identifies and localizes safety neurons that are associated with safe and reliable model behavior. It then applies gradient masking to irrelevant neurons, restricting parameter updates to a safety subspace and thereby enabling precise, targeted safety fine-tuning with only a very small fraction of trainable parameters. This design substantially improves model safety while largely preserving the original multilingual, multimodal, and general-purpose capabilities of the underlying models. Extensive experiments on three state-of-the-art VLLMs further demonstrate the effectiveness, targeting, and generalizability of the proposed approach across different model architectures, languages, risk modalities, and jailbreak scenarios.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661* (2025).

[2] Anjanava Biswas and Wrick Talukdar. 2026. Guardrails for trust, safety, and ethical development and deployment of Large Language Models (LLM). *arXiv preprint arXiv:2601.14298* (2026).

[3] Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. 2026. Align Once, Benefit Multilingually: Enforcing Multilingual Consistency for LLM Safety Alignment. In *The Fourteenth International Conference on Learning Representations*.

[4] Zouying Cao, Yifei Yang, and Hai Zhao. 2025. SCANS: Mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 23523–23531.

[5] Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Qizhe Shieh, and Wenxuan Zhang. 2025. The Emergence of Abstract Thought in Large Language Models Beyond Any Language. *arXiv preprint arXiv:2506.09890* (2025).

[6] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations*, Vol. 2024. 24634–24651.

[7] Yi Ding, Bolian Li, and Ruqi Zhang. 2024. Eta: Evaluating then aligning safety of vision language models at inference time. *arXiv preprint arXiv:2410.06625* (2024).

[8] Jingtong Dou, Chuancheng Shi, Yemin Wang, Shiming Guo, Anqi Yi, Wenhua Wu, Li Zhang, Fei Shen, and Tat-Seng Chua. 2026. DNA: Uncovering Universal Latent Forgery Knowledge. *arXiv preprint arXiv:2601.22515* (2026).

[9] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355* (2024).

[10] Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025. H-Neurons: On the Existence, Impact, and Origin of Hallucination-Associated Neurons in LLMs. *arXiv preprint arXiv:2512.01797* (2025).

[11] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 23951–23959.

[12] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*. Springer, 388–404.

[13] 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* (2024).

[14] Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. 2024. Cold-attack: Jailbreaking llms with stealthiness and controllability. *arXiv preprint arXiv:2402.08679* (2024).

[15] Bing Han, Feifei Zhao, Dongcheng Zhao, Guobin Shen, Ping Wu, Yu Shi, and Yi Zeng. 2025. Fine-Grained Safety Neurons with Training-Free Continual Projection to Reduce LLM Fine Tuning Risks. *arXiv preprint arXiv:2508.09190* (2025).

[16] Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2025. The emerged security and privacy of llm agent: A survey with case studies. *Comput. Surveys* 58, 6 (2025), 1–36.

[17] Raviraj Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Long, Sanjay Singh Chauhan, et al. 2025. CultureGuard: Towards Culturally-Aware Dataset and Guard Model for Multilingual Safety Applications. *arXiv preprint arXiv:2508.01710* (2025).

[18] Nomisha Kurian. 2025. 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap'in large language models. *Learning, Media and Technology* 50, 4 (2025), 621–634.

[19] Jiaming Liang, Zhaoxin Wang, and Handing Wang. 2026. Multilingual Safety Alignment Via Sparse Weight Editing. *arXiv preprint arXiv:2602.22554* (2026).

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

[21] Siyuan Ma, Weidi Luo, Yu Wang, and Xiaogeng Liu. 2024. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773* (2024).

[22] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.

[23] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309* (2024).

[24] Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. [n. d.]. LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked. In *The Second Tiny Papers Track at ICLR 2024*.

[25] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 5377–5400.

[26] Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136* (2024).

[27] Enyi Shi, Pengyang Shao, Yanxin Zhang, Chenhang Cui, Jiayi Lyu, Xu Xie, Xiaobo Xia, Fei Shen, and Tat-Seng Chua. 2026. Lingua-SafetyBench: A Benchmark for Safety Evaluation of Multilingual Vision-Language Models. *arXiv preprint arXiv:2601.22737* (2026).

[28] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057* (2022).

[29] 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. *arXiv preprint arXiv:2402.16438* (2024).

[30] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[31] Han Wang, Gang Wang, and Huan Zhang. 2025. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29947–29957.

[32] Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*. 5865–5877.

[33] Zhaoxin Wang, Jiaming Liang, Fengbin Zhu, Weixiang Zhao, Junfeng Fang, Jiayi Ji, Handing Wang, and Tat-Seng Chua. 2026. SafeNeuron: Neuron-Level Safety Alignment for Large Language Models. *arXiv preprint arXiv:2602.12158* (2026).

[34] 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* (2025).

[35] Xin Yi, Shunfan Zheng, Linlin Wang, Gerard de Melo, Xiaoling Wang, and Liang He. 2025. Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 25706–25714.

[36] Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. 2025. Jailbreak vision language models via bi-modal adversarial prompt. *IEEE Transactions on Information Forensics and Security* (2025).

[37] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6216–6226.

[38] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* (2023).

[39] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. 2026. Instruction tuning for large language models: A survey. *Comput. Surveys* 58, 7 (2026), 1–36.

[40] Xianhui Zhang, Chengyu Xie, Linxia Zhu, Yonghui Yang, Weixiang Zhao, Zifeng Cheng, Cong Wang, Fei Shen, and Tat-Seng Chua. 2026. Who Transfers Safety? Identifying and Targeting Cross-Lingual Shared Safety Neurons. *arXiv preprint arXiv:2602.01283* (2026).

[41] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. 2025. Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19867–19878.

[42] Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. 2025. Qwen3Guard Technical Report. *arXiv preprint arXiv:2510.14276* (2025).

[43] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).

Appendix

The appendices provide additional details that support and extend the main paper. Appendix A describes the datasets used in our experiments. Appendix B provides further experimental details and analyses of the results. Appendix C presents a more in-depth theoretical analysis of the proposed our method. Appendix D intuitively demonstrates the effectiveness of our method through layer-wise feature-space comparisons with the vanilla model and additional qualitative examples. Appendix E offers a more in-depth discussion of our method. Finally, Appendix F summarizes the limitations of this work and outlines directions for future research.

A Dataset Overview

In our experiments, we incorporate Lingua-SafetyBench [27], MM-Bench [20], MM-Vet [38], MGSM [28], XSTest [25], FigStep [11], SPA-VL [41], MultiJail [6] and AdvBench [43], each serving a distinct role in neuron probing, capability evaluation, over-refusal analysis, or out-of-distribution safety testing.

Lingua-SafetyBench [27] We use Lingua-SafetyBench as the primary benchmark to investigate safety neuron activation patterns in VLLMs. It supports multilingual safety evaluation and distinguishes Image-Dominant Risk, where harmful semantics are conveyed through visual content, from Text-Dominant Risk, where harmful intent is expressed in text. The benchmark covers a diverse range of harmful scenarios. We curate a total of 14,910 challenging samples across 10 languages and divide them into two disjoint subsets. Specifically, 9,830 samples are used for neuron probing and safety tuning, and all training-based baselines are trained on the same subset to ensure a fair comparison. The remaining 5,080 samples are used for safety evaluation. This protocol prevents data overlap across different stages and enables a reliable assessment of multilingual and multimodal safety alignment.

MMBench [20] MMBench is a multimodal benchmark for objectively evaluating vision-language large models, covering a wide range of capability dimensions including perception, reasoning, and knowledge understanding. Building upon this benchmark, we analyze neuron activation patterns on MMBench to characterize how VLLMs respond to general multimodal requests. Based on the previously identified safety neurons, we further use the multilingual version of MMBench and sample 1,000 benign evaluation instances per language to identify general-capability neurons, which are then removed from the safety-neuron set. This yields a more specific set of safety neurons while minimizing interference with the model's general multimodal capabilities.

MM-Vet [38] MM-Vet evaluates large multimodal models on complex tasks by assessing their ability to integrate multiple vision-language capabilities. It defines six core capabilities and sixteen combinations, covering tasks such as visual understanding and cross-modal reasoning. We use MM-Vet to study the impact of safety alignment on general multimodal capabilities by comparing performance before and after alignment.

MGSM [28] MGSM is a benchmark for evaluating multilingual reasoning abilities using math word problems that require multi-step reasoning across diverse languages. We use MGSM to measure multilingual reasoning capabilities and analyze how safety alignment affects overall multilingual performance.

XSTest [25] XSTest is a benchmark specifically designed to evaluate over-refusal. It contains diverse benign prompts that should not be refused, enabling systematic assessment of unnecessary safety refusals. We use the benign samples in XSTest to measure both refusal and compliance rates, thereby evaluating whether safety alignment causes excessive refusal of harmless user requests.

FigStep [11] FigStep is a multimodal jailbreak benchmark that converts harmful textual instructions into typographic images to bypass the safety alignment of vision-language large models. It probes whether models can reliably recognize unsafe intent when malicious content is conveyed through the visual channel. We use FigStep to evaluate safety robustness under image-based jailbreak attacks as a challenging out-of-distribution scenario.

SPA-VL [41] SPA-VL is a large-scale multimodal harmful-request dataset for vision-language large models, covering 6 harmfulness domains, 13 categories, and 53 subcategories. Each sample consists of a harmful question and a corresponding image. We use SPA-VL as an out-of-distribution test set to evaluate safety robustness under multimodal harmful requests.

MultiJail [6] MultiJail is a multilingual jailbreak benchmark for robustly evaluating safety vulnerabilities of language models across languages. It covers both unintentional and intentional multilingual jailbreak scenarios, where unsafe intent may be expressed or amplified through non-English prompts. We use MultiJail to effectively evaluate safety generalization under multilingual OOD scenarios involving harmful intent across diverse languages.

AdvBench [43] AdvBench is a harmful-instruction benchmark for evaluating the safety robustness of large language models. We use representative attacks built upon AdvBench, including GCG, ImgJP, and BAP, as out-of-distribution test sets to evaluate safety robustness against open-world adversarial attacks.

B More Details and Results

Implementation Details. To facilitate reproducibility, we provide comprehensive implementation details, including the training configurations and hyperparameter settings, as summarized in Table 8. We also provide the complete safety-aware prompt used during the neuron probing stage to guide VLLMs toward generating refusal responses, as presented in Prompt 1. For safety evaluation, we conduct automated safety assessment using Qwen-Guard and Llama-Guard. Since both are specialized safety guard models with officially specified inference and evaluation protocols, we do not introduce any manually engineered judging prompts and instead strictly follow their official default configurations.

Prompt 1: Safety-Aware Refusal Prompt

The following multimodal input contains confirmed harmful content, either in the image or text. When analyzing the multimodal input, you must refuse the request and briefly explain the reason, unless explicitly instructed to provide an extremely safe response. The explanation itself must not contain any harmful content.

Sensitivity Analysis. To investigate the sensitivity of the intervention ratio p, we conduct additional experiments on Qwen using four representative settings within the range of 0.01–0.15. As shown

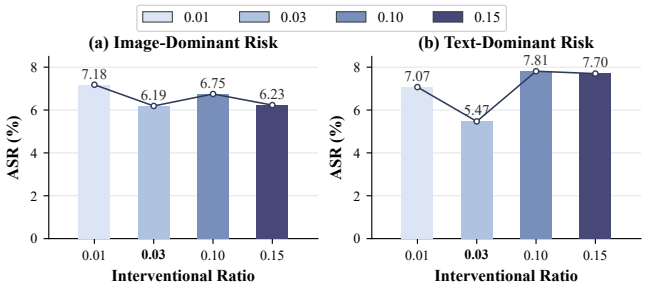


Figure 10: Sensitivity analysis of the intervention ratio. Our method remains relatively robust across values, while the selected (0.03) achieves the best safety performance.

in Figure 10, our method maintains consistently strong safety performance across different intervention ratios, demonstrating its robustness to the choice of p. Among all settings, $p = 0.03$ achieves the best overall safety performance, with the lowest average ASRs of 6.19 and 5.47 on Image-Dominant Risk and Text-Dominant Risk, respectively, across all 10 languages. This result further suggests that strong multilingual and multimodal safety alignment can be achieved by intervening in only a small proportion of safety neurons. We therefore adopt $p = 0.03$ as the default setting.

LLM Judge Reliability. To assess the reliability of Qwen-Guard as an automated judge, we randomly select 1,280 samples through cross-lingual stratified balanced sampling and manually annotate their safety labels. As shown in Table 9, Qwen-Guard achieves consistently high judgment accuracy across all languages, with an average accuracy of 92.89%. These results demonstrate its strong agreement with human annotations across diverse linguistic settings, validating its use as a reliable evaluator in our experiments.

C Theoretical Analysis

We provide an in-depth theoretical analysis of our method. Let $\boldsymbol{\theta}$ collect the vectorized entries of the LoRA $\boldsymbol{B}$ matrices. To isolate the effect of neuron-targeted masking, we hold the remaining parameter blocks fixed during the analyzed step. We denote the safety-alignment loss and the general-capability loss by $\mathcal{L}_{\mathrm{s}}$ and $\mathcal{L}_{\mathrm{u}}$, respectively. Because the binary neuron mask acts coordinate-wise on $\boldsymbol{B}$, it induces a diagonal orthogonal projector $\boldsymbol{P}$ satisfying $\boldsymbol{P}^{\top} = \boldsymbol{P}$ and $\boldsymbol{P}^2 = \boldsymbol{P}$. This projector restricts optimization to the selected coordinates while leaving all complementary coordinates unchanged, thereby separating the choice of an intervention subspace from the magnitude of the update performed within it. At the current iterate, we formally define

$$\boldsymbol{g}_{\mathrm{s}} := \nabla\mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}), \boldsymbol{g}_{\mathrm{u}} := \nabla\mathcal{L}_{\mathrm{u}}(\boldsymbol{\theta}), \boldsymbol{d} := \boldsymbol{P}\boldsymbol{g}_{\mathrm{s}}, \boldsymbol{\theta}^{+} := \boldsymbol{\theta} - \eta\boldsymbol{d}, \tag{13}$$

where $\eta > 0$ is the step size. Assume that $\mathcal{L}_{\mathrm{s}}$ and $\mathcal{L}_{\mathrm{u}}$ have β_{s}- and β_{u}-Lipschitz gradients, respectively. We further characterize the selected subspace using a gradient-retention coefficient $\kappa \in (0, 1]$ and an interference coefficient $\rho \in [0, 1]$:

$$\|\nabla\mathcal{L}_a(\boldsymbol{x}) - \nabla\mathcal{L}_a(\boldsymbol{y})\|_2 \le \beta_a\|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad a \in \{\mathrm{s}, \mathrm{u}\}, \tag{14}$$

$$\|\boldsymbol{d}\|_2 \ge \kappa\|\boldsymbol{g}_{\mathrm{s}}\|_2, \qquad \left|\langle\boldsymbol{g}_{\mathrm{u}}, \boldsymbol{d}\rangle\right| \le \rho\|\boldsymbol{g}_{\mathrm{u}}\|_2\|\boldsymbol{d}\|_2. \tag{15}$$

Here, κ quantifies how much of the safety gradient remains available after masking, whereas ρ measures the first-order coupling between the masked safety direction and the utility objective. These coefficients describe two distinct properties of the selected neurons. A large κ ensures that projection does not discard the direction needed to reduce the safety loss. A small ρ ensures that the retained direction is nearly orthogonal to the utility gradient and therefore has little first-order influence on general capabilities. Neither condition alone is sufficient: retaining a strong safety direction may still harm utility, while eliminating utility interference may leave too little safety signal for an effective update.

Table 8: Implementation details across VLLMs.

Configuration	LLaVA	Llama	Qwen
Computing Device	1× A800	1× A800	1× A800
Global Batch Size	32	32	32
Training Epochs	5	5	5
Optimizer	AdamW	AdamW	AdamW
Learning Rate	6e−4	6e−4	6e−4
Warmup Ratio	0.05	0.05	0.05
Weight Decay	0.0	0.0	0.0
Dropout	0.05	0.05	0.05
Random Seed	42	42	42
Rank	8	8	8
Alpha	16	16	16
Intervention Ratio	0.03	0.03	0.03

Table 9: Qwen-Guard Judgment Accuracy (%). Qwen-Guard demonstrates strong and consistent alignment with human judgments across 10 diverse languages.

Language	Arabic	Chinese	English	Finnish	French
Accuracy (%)	91.41	92.19	90.63	95.31	96.09
Language	**German**	**Japanese**	**Norwegian**	**Russian**	**Spanish**
Accuracy (%)	92.97	94.53	90.63	94.53	90.63
			Average Accuracy (%)		**92.89**

Proposition C.1 (Safety descent and utility perturbation). *Suppose that Equations (14) and (15) hold. For any $0 < \eta \le 1/\beta_{\mathrm{s}}$, the update in Equation (13) satisfies*

$$\mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}^{+}) \ge \eta\left(1 - \frac{\beta_{\mathrm{s}}\eta}{2}\right)\|\boldsymbol{d}\|_2^2 \ge \frac{\eta\kappa^2}{2}\|\boldsymbol{g}_{\mathrm{s}}\|_2^2, \tag{16}$$

$$\left|\mathcal{L}_{\mathrm{u}}(\boldsymbol{\theta}^{+}) - \mathcal{L}_{\mathrm{u}}(\boldsymbol{\theta})\right| \le \eta\rho\|\boldsymbol{g}_{\mathrm{u}}\|_2\|\boldsymbol{d}\|_2 + \frac{\beta_{\mathrm{u}}\eta^2}{2}\|\boldsymbol{d}\|_2^2. \tag{17}$$

Consequently, the safety loss is non-increasing and decreases strictly whenever $\boldsymbol{d} \neq \boldsymbol{0}$.

The two inequalities expose an important asymmetry. The safety improvement is of order η, whereas the utility perturbation contains a first-order interference term scaled by ρ and a second-order curvature term. Thus, a small ρ suppresses the dominant utility change, while a sufficiently large κ preserves nontrivial safety descent.

By β_{s}-smoothness and $\boldsymbol{\theta}^{+} - \boldsymbol{\theta} = -\eta\boldsymbol{d}$, we have

$$\begin{aligned}\mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}^{+}) &\le \mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}) - \eta\langle\boldsymbol{g}_{\mathrm{s}}, \boldsymbol{d}\rangle + \frac{\beta_{\mathrm{s}}\eta^2}{2}\|\boldsymbol{d}\|_2^2 \\ &= \mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}) - \eta\left(1 - \frac{\beta_{\mathrm{s}}\eta}{2}\right)\|\boldsymbol{d}\|_2^2.\end{aligned} \tag{18}$$



Figure 11: Image-Dominant Risk visualization cases across multiple languages.

The equality follows from $\langle \boldsymbol{g}_s, \boldsymbol{P}\boldsymbol{g}_s \rangle = \|\boldsymbol{P}\boldsymbol{g}_s\|_2^2$, since $\boldsymbol{P}$ is an orthogonal projector. Because $\eta \le 1/\beta_s$, we have $1 - \beta_s\eta/2 \ge 1/2$. Combining this inequality with $\|\boldsymbol{d}\|_2 \ge \kappa\|\boldsymbol{g}_s\|_2$ proves Equation (16). For the utility objective, smoothness gives

$$\left|\mathcal{L}_u(\boldsymbol{\theta}^+) - \mathcal{L}_u(\boldsymbol{\theta}) + \eta \left\langle \boldsymbol{g}_u, \boldsymbol{d} \right\rangle\right| \le \frac{\beta_u \eta^2}{2}\|\boldsymbol{d}\|_2^2. \tag{19}$$

Applying the triangle inequality and the interference condition in Equation (15) proves Equation (17).

The coordinate projector also makes the sparsity of the update explicit. If $\mathrm{rank}(\boldsymbol{P}) = m$, where m is the number of active entries in the masked $\boldsymbol{B}$ matrices, then $\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_0 \le m$ and $\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_2 = \eta\|\boldsymbol{d}\|_2$. Hence, the rank of $\boldsymbol{P}$ directly controls the extent of modification, independently of the ambient parameter dimension.

Moreover, when $\rho = 0$ and $\boldsymbol{d} \ne \boldsymbol{0}$, Equations (16) and (17) imply

$$\frac{\left|\mathcal{L}_u(\boldsymbol{\theta}^+) - \mathcal{L}_u(\boldsymbol{\theta})\right|}{\mathcal{L}_s(\boldsymbol{\theta}) - \mathcal{L}_s(\boldsymbol{\theta}^+)} \le \frac{\beta_u \eta}{2 - \beta_s \eta} \le \beta_u \eta. \tag{20}$$

Thus, in the absence of first-order interference, the improvement is first order in η, whereas the utility variation is second order.

The theoretical result shows that if the masked update direction retains sufficient safety-gradient information, corresponding to a large κ, while its interference with the general-capability gradient remains small, corresponding to a small ρ, then a neuron-targeted update is guaranteed to decrease the safety loss while keeping the change in the utility loss within an explicit upper bound. In addition, row-wise masking restricts the set of modified parameters, making the update sparse. The analysis proves that, when an effective safety direction is preserved and capability interference is suppressed, targeted updates can simultaneously improve safety, control utility degradation, and limit the extent of parameter modification.

D Visualization and Cases Analysis

Qualitative Case Visualization. To more intuitively demonstrate the effectiveness of our method, we present additional qualitative examples in Figure 11 and Figure 12, covering both Image-Dominant and Text-Dominant Risk scenarios across multiple languages and three representative VLLMs, including Llama, LLaVA, and Qwen. In these cases, the vanilla models are easily bypassed by multilingual–multimodal harmful requests and consequently generate unsafe responses. In contrast, our method consistently recognizes the underlying harmful intent and produces appropriate safety refusals. This improvement is attributed to the targeted reinforcement



Figure 12: Text-Dominant Risk visualization cases across multiple languages.

of safety neurons, which strengthens the models' ability to identify and mitigate risks across different languages and modalities. These qualitative examples further demonstrate the effectiveness and generalizability of our method in defending against diverse multilingual–multimodal attacks.

Progressive Safety Intervention. To better illustrate how our method affects the internal representations of vision-language large models (VLLMs) after the targeted reinforcement of safety neurons, we employ t-SNE [30] to visualize layer-wise feature distributions. Specifically, we select multimodal harmful requests in low-resource languages that induce the vanilla model to generate harmful responses, covering both Image-Dominant Risk and Text-Dominant Risk settings. We then compare the layer-wise representations produced by our method and the vanilla model when processing these inputs, as shown in Figures 13 and 14. Although the feature representations of the two models remain relatively close and partially overlap in the shallow layers, the representations produced by our method become clearly separated from those of the vanilla model in the middle and deeper layers. Notably, this separation emerges earlier under Text-Dominant Risk, at approximately Layer 8, whereas under Image-Dominant Risk, a clear intervention effect appears only around Layer 12. This difference suggests that harmful intent conveyed directly through linguistic cues can be recognized and redirected by the reinforced safety neurons at an earlier stage of representation processing. In contrast, Image-Dominant Risk requires more complex cross-modal grounding and visual-semantic integration before the harmful intent can be fully identified, causing the effect of safety-neuron intervention to emerge in deeper layers. These observations indicate that the targeted reinforcement of safety neurons progressively alters the representation trajectories of harmful inputs during deeper computation, steering the model's internal representations toward a safer direction and thereby leading to safe responses.

E More Discussion

▹ ***Q1. The proposed method updates only a very small fraction of parameters (< 0.03%). Could this sparsity limit generalization to unseen or out-of-distribution (OOD) attack types, compared to full fine-tuning or LoRA?***

A: We argue that this sparsity is a strength rather than a limitation. Although our method updates less than 0.03% of the model parameters, it targets a compact set of neurons that are functionally associated with safety-related behaviors instead of modifying

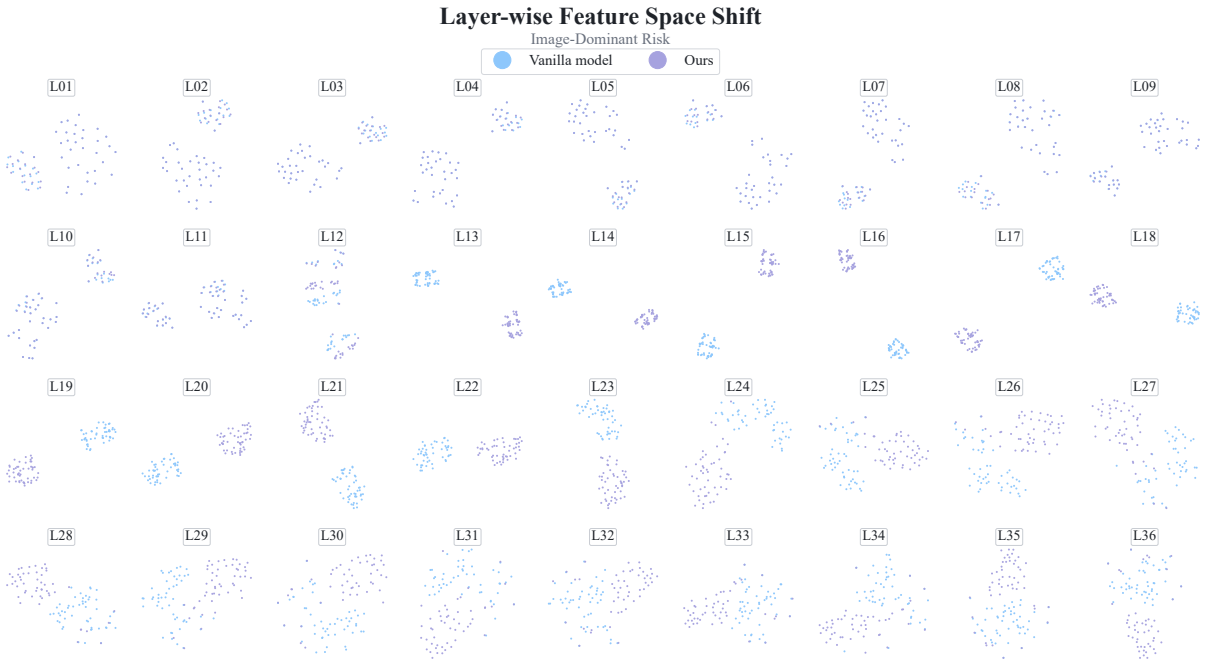


Figure 13: Layer-wise feature space shifts under Image-Dominant Risk. Our method gradually separates harmful-input representations from those of the vanilla model in deeper layers, steering them toward a safer representation space.

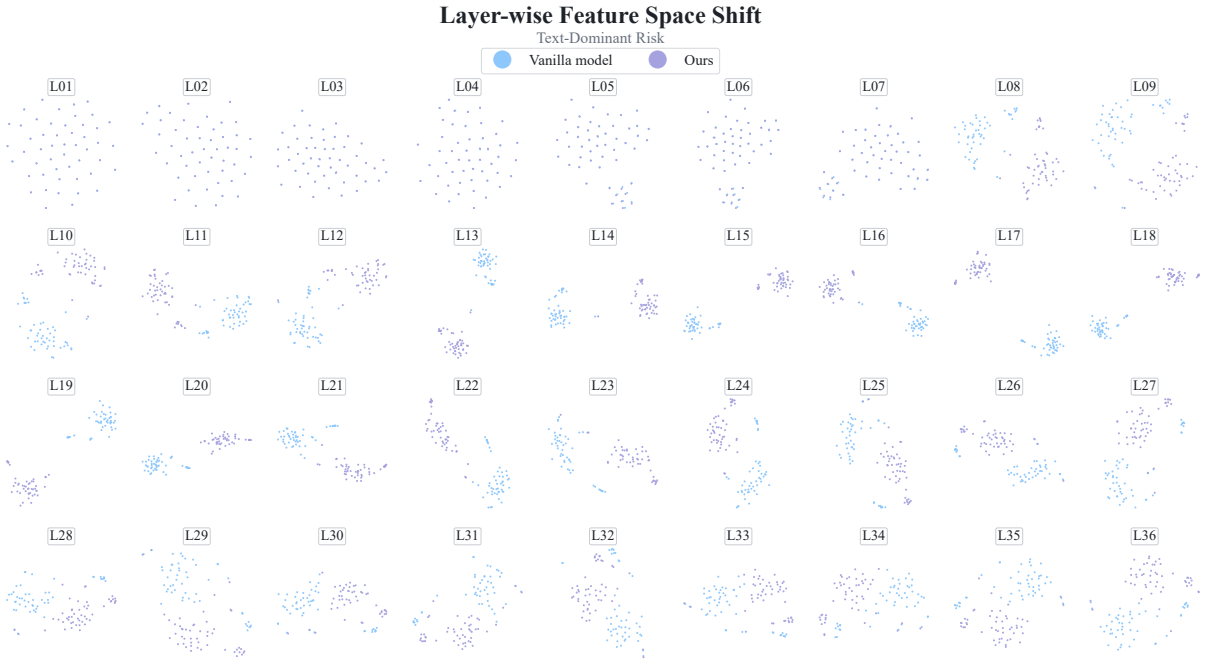


Figure 14: Layer-wise feature space shifts under Text-Dominant Risk. Our method induces an earlier and progressively clearer separation of harmful-input representations across layers, steering them toward a safer representation space.

the model indiscriminately. We systematically evaluate its generalization under diverse out-of-distribution settings, including the multimodal safety datasets SPA-VL and FigStep, the multilingual jailbreak dataset MultiJail, and representative attack methods such

as GCG, ImgJP, and BAP. Across these unseen datasets, languages, modalities, and attack settings, our method consistently reduces ASR, demonstrating that its effectiveness extends beyond the training distribution. These results indicate that sparse updates to safety neurons do not compromise OOD generalization. Instead, by focusing optimization on the safety-relevant functional subspace, our method strengthens transferable safety representations while avoiding unnecessary parameter perturbations, thereby reducing the risk of catastrophic forgetting compared with less selective parameter-efficient methods such as LoRA.

▹ ***Q2. The safety neuron selection involves excluding general-purpose neurons via a set difference operation. Could this process inadvertently remove polysemantic neurons that contribute both to safety and general multimodal understanding?***

A: This is an important consideration regarding the balance between safety and utility. Our strategy is designed to prioritize the preservation of general capabilities by filtering out neurons that are consistently activated in benign multimodal tasks from MMBench. While it is possible that some polysemantic neurons contributing to both safety and general understanding are inadvertently excluded, empirical results on MM-Vet and MGSM show that general performance is well preserved and, in some cases, even slightly improved. This suggests that the remaining safety neurons are sufficient to support robust alignment, and that excluding shared neurons is an effective strategy for avoiding unintended degradation of core multilingual and multimodal abilities. These results further indicate that our overall set-difference strategy effectively balances robust safety alignment with the preservation of core general multimodal and multilingual capabilities.

▹ ***Q3. The analysis focuses primarily on FFN layers for neuron probing. Could you elaborate on the rationale for not considering Attention layers, given their role in cross-modal information integration in Transformer architectures?***

A: Our focus on FFN layers is motivated by the knowledge-based view of Transformers, in which FFNs are often regarded as key components for storing and transforming semantic knowledge. While Attention layers are crucial for regulating information flow and facilitating cross-modal interactions, their patterns are typically distributed across tokens, heads, and modalities, making neuron-level attribution less direct. In contrast, FFN activations provide a more localized and interpretable signal of how the model responds to semantic triggers such as harmful intent. This makes FFNs particularly suitable for identifying safety-relevant neurons and performing precise parameter interventions. Importantly, our empirical results show that modifying less than 0.03% of FFN parameters substantially improves safety across languages, modalities, and attack types while preserving general capabilities. These findings suggest that FFN layers constitute an important locus of safety-related representations, although jointly analyzing Attention and FFN components remains a valuable direction for future work.

▹ ***Q4. The analysis shows a moderate overlap of safety neurons across languages. Does this suggest that the method may need to be re-applied for each newly introduced target language, potentially increasing deployment complexity?***

A: Although the overlap of safety neurons across languages is moderate rather than complete, it remains statistically meaningful and reveals the existence of a shared internal safety subspace. Our zero-shot transfer analysis further shows that intervening on safety neurons identified in one language can improve safety performance in other languages without requiring additional neuron probing or retraining. This indicates that the identified neurons capture both language-specific and language-agnostic safety features. In practical deployment, one can first construct a transferable core safety map using several high-resource or representative languages and directly apply it to newly introduced low-resource languages. Additional language-specific probing would only be necessary when stronger adaptation is desired. Therefore, our method does not require exhaustive per-language recomputation and can support scalable multilingual deployment. The moderate overlap also provides flexibility, allowing the framework to combine a shared safety backbone with lightweight language-specific enhancement when needed.

▹ ***Q5. Why is neuron importance measured by jointly combining activation magnitude with the norm of the corresponding down-projection vector, rather than using activation alone or gradient-based attribution methods?***

A: Activation magnitude alone reflects only how strongly a neuron responds to a given input, but does not indicate how much it influences the subsequent hidden representation. Conversely, the norm of the corresponding column in $\mathbf{W}_{\text{down}}$ characterizes the neuron's potential downstream influence, but does not capture whether it is actually activated by harmful inputs. Their product therefore jointly measures input-dependent responsiveness and effective downstream impact in the model, providing a more complete estimate of a neuron's functional contribution than either factor alone. Compared with gradient-based attribution methods, this forward-based metric is independent of a specific training loss or target response, avoids potentially noisy and unstable backward signals, and enables consistent neuron comparison across different languages, modalities, and datasets. Moreover, the neuron-masking and targeted-tuning experiments further verify that the neurons selected by this score are functionally relevant to model safety rather than merely being highly activated general-purpose neurons.

F Limitations and Future Work

While our method demonstrates performance and provides mechanistic insights into VLLM safety, several limitations remain, motivating broader evaluation across diverse real-world deployment contexts. First, it focuses on feed-forward network layers and may overlook interactions with attention modules, residual pathways, and cross-modal projection layers. Future work could examine how safety signals propagate across the architecture for a comprehensive understanding of multimodal safety alignment. Second, our evaluation relies on existing datasets, which may not capture regional dialects, culturally specific expressions, or context-dependent safety concerns. Culture-aware and context-sensitive benchmarks are needed to assess generalizability more broadly in diverse settings. Finally, our method relies on static safety definitions, which may limit adaptability as human values and ethical norms evolve. Incorporating dynamic value systems could support adaptive long-term alignment while preserving interpretability.