

# Finding Culture-Sensitive Neurons in Vision-Language Models

Xiutian Zhao<sup>1,3</sup> Rochelle Choenni<sup>2</sup> Rohit Saxena<sup>1</sup> Ivan Titov<sup>1,2</sup>

<sup>1</sup>University of Edinburgh <sup>2</sup>University of Amsterdam <sup>3</sup>Johns Hopkins University

## Abstract

Despite their impressive performance, vision-language models (VLMs) still struggle on culturally situated inputs. To understand how VLMs process culturally grounded information, we study the presence of culture-sensitive neurons, i.e., neurons whose activations show preferential sensitivity to inputs associated with particular cultural contexts. We examine whether such neurons are important for culturally diverse visual question answering and where they are located. Using the CVQA benchmark, we identify neurons of culture selectivity and perform diagnostic tests by deactivating the neurons flagged by various identification methods. Experiments on three VLMs across 25 cultural groups demonstrate the existence of neurons whose ablation disproportionately harms performance on questions about the corresponding cultures, while having limited effects on others. Moreover, we introduce a new margin-based selector—Contrastive Activation Margin (ConAct)—and show that it outperforms probability- and entropy-based methods in identifying neurons associated with cultural selectivity. Finally, our layer-wise analyses reveal that such neurons are not uniformly distributed: they cluster in specific decoder layers in a model-dependent way. <sup>1</sup>

## 1 Introduction

Vision-language models (VLMs) underpin many multimodal applications, from visual question answering (VQA) to chart captioning and document parsing (Liu et al., 2023; Li et al., 2023; Bai et al., 2025; Yue et al., 2025). Despite impressive performance, various works show that many VLMs struggle on culturally grounded visual content or culturally marked linguistic cues, and often exhibit systematic performance disparities across cultures (Romero et al., 2024; Nayak et al., 2024). Understanding how and where such culture-related

knowledge is represented within VLMs is important both for interpretability and fairness. Identifying subcomponents that are important for culture-related processing can not only improve our understanding of the underlying mechanisms, but may also guide future efforts to enhance these capabilities during post-training, e.g., through sparse fine-tuning (Ansell et al., 2022; Ben Zaken et al., 2022) or activation steering (Turner et al., 2024; Rimsky et al., 2024).

Prior work in neural network interpretability has shown that individual neurons can exhibit relative specialization for certain concepts, modalities, or tasks (Bau et al., 2017, 2020). In large language models (LLMs), researchers have found neurons that are preferentially active for particular languages (Tang et al., 2024), knowledge domains (Yu and Ananiadou, 2024) and text-styles (Lai et al., 2024). Analyses of VLMs, however, have primarily focused on modality-related aspects when identifying neuron functions (e.g., distinguishing neurons involved in visual vs. textual processing) (Huang et al., 2024; Fang et al., 2024; Xu et al., 2025), leaving other forms of specialization underexplored. Specifically, it is unknown whether VLMs contain neurons that preferentially respond to inputs from specific cultural contexts, as opposed to comparable inputs from others. This question is especially relevant given that culture-related signals often arise from interactions between the visual and textual modalities. Addressing this gap can shed new light on how VLMs encode culturally grounded knowledge and where possible limitations or biases originate.

Thus, we study whether VLMs contain neurons whose activity is selectively modulated by culturally grounded inputs, without implying that these neurons are exclusively dedicated to culture. Instead, we aim to identify neurons that show relative culture-selectivity, i.e., units whose activations exhibit stronger association with certain cultural con-

<sup>1</sup>Related code is available at <https://github.com/xiutian/vlm-culture-neuron>.

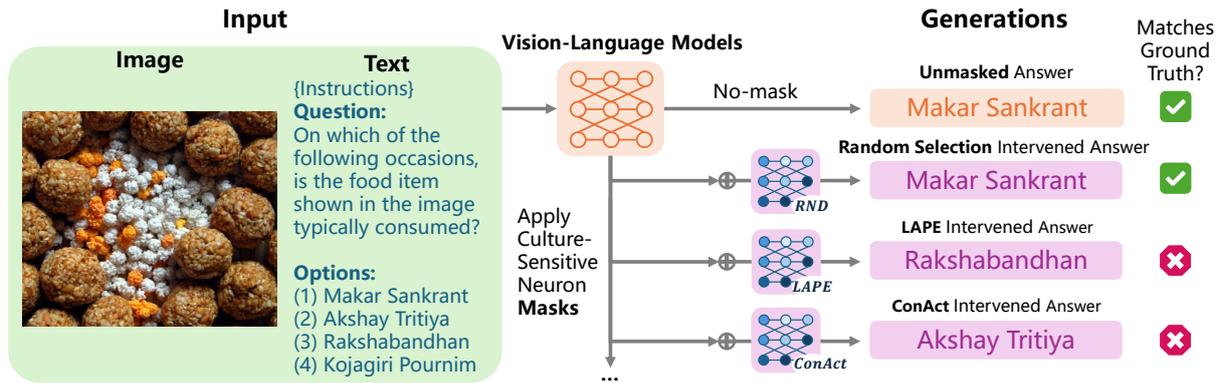


Figure 1: **An ablation example** of Qwen2.5-VL-7B on India-Marathi VQA subset. Given an image of Tilgul, an Indian sweet made from sesame seeds and jaggery, the full model selects the ground truth-matched option; **RND** mask does not affect the model’s decision, while **LAPE** and **ConAct** masks redirect to different answers. Mentioned methods are explained in § 3.2.

texts compared to others, and to evaluate to what extent such neurons are critical for culture-specific performance. Concretely, we address the following questions: (1) Do VLMs contain such culture-sensitive neurons, i.e., neurons that preferentially activate on inputs tied to particular cultures? (2) Does ablating small, targeted subsets of these neurons selectively degrade a VLM’s performance on questions tied to the corresponding culture, with minimal impact on other cultures? (3) How are these neurons distributed across layers, and is the pattern consistent across model architectures and cultures?

Following prior work on neuron detection (Tang et al., 2024; Huo et al., 2024; Huang et al., 2024; Fang et al., 2024), we adapt activation-based neuron analysis to a multimodal setting and evaluate on the CVQA benchmark (Romero et al., 2024), operationalizing culture via the CVQA taxonomy of country–language pairs. We conduct experiments on three VLMs : Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-v1.6-Mistral-7B (Liu et al., 2023), and Pangea-7B (Yue et al., 2025), across 25 cultures. To minimize influence from differences in language proficiency or language-correlated effects, we constrain the experiments to a monolingual (English) setting. Moreover, to better isolate culture-sensitive neurons, we introduce *Contrastive Activation Margin* (ConAct), a margin-based method that rewards large separation between a neuron’s activation for its top-responding culture and its nearest competing culture, improving upon existing probability- and entropy-based selectors.

We provide empirical evidence for the existence of culture-sensitive neurons in VLMs. Ablating

these neurons disproportionately reduces model performance on questions tied to the corresponding culture while leaving others largely unaffected, suggesting a causal role in culturally grounded information processing. Moreover, our layer-wise analysis reveals that these neurons are distributed across the decoder, with noticeable concentrations in mid-to-late layers. While we do observe some exceptions, this pattern remains largely consistent across the VLMs and cultures we examine. Overall, our results provide insight into how VLMs represent cultural knowledge and suggest new avenues for targeted evaluation and intervention to mitigate cultural biases or steer model behavior.

## 2 Related work

**Studying neuron specialization.** Identifying specialized neurons that respond strongly to particular features or concepts is a well-established practice in interpreting deep neural network models. Early work on convolutional neural network interpretability (Bau et al., 2017, 2020) showed that individual hidden units can align with human-understandable concepts, such as objects, parts, colors, or even high-level concepts. Analogous analyses have been applied to modern LLMs. For instance, Yu and Ananiadou (2024) showed potential neurons specialized at domain-knowledge; Tang et al. (2024) introduced an entropy-based method to find language-specific neurons. Two relevant works demonstrated evidence of culture-related neurons in LLMs (Namazifard and Poeh, 2025; Yamamoto et al., 2025). However, in multimodal settings, existing efforts are concentrated on identifying modality- (Pan et al., 2024; Huang

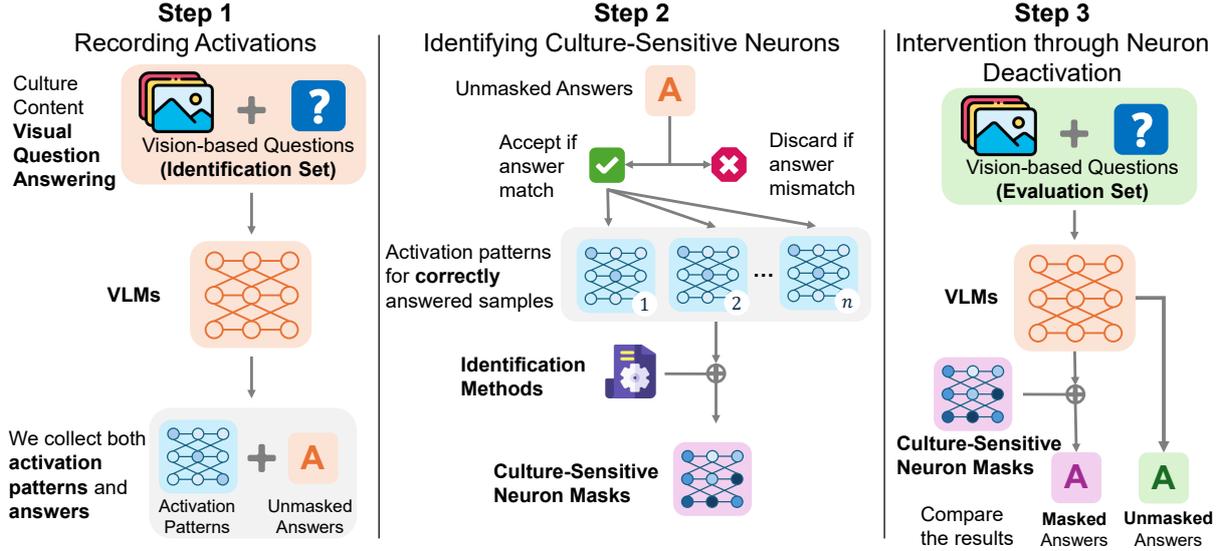


Figure 2: Pipeline for identifying and validating culture-sensitive neurons: (1) record neuron activations on culture-specific VQAs, (2) identify influential neurons using several methods, and (3) evaluate their importance by ablating the top- $r$ % neurons and measuring the effect on accuracy and answer divergence.

et al., 2024; Fang et al., 2024; Xu et al., 2025) or task-specific neurons (Neo et al., 2025), whereas domain-knowledge associated neurons are limited (Huo et al., 2024; Zhao et al., 2026).

**Cultural knowledge and bias in VLMs.** Culture is a complex, multifaceted construct involving shared knowledge, practices, symbols and social norms of a group (Tylor, 1871; Hofstede, 1980). Culture-related multimodal benchmarks, such as CVQA (Romero et al., 2024), CULTURALVQA (Nayak et al., 2024), and CULTURALGROUND (Nyandwi et al., 2025), approximate culture via local knowledge and practices that are common in a region or within a language group (Pawar et al., 2025). Such datasets test VLMs on culturally diverse content (e.g., traditional foods, clothing, landmarks), often revealing VLMs’ substantial performance disparities across different cultures. Moreover, prior studies have found that VLMs tend to exhibit systematic biases both in the image perception and natural language reasoning (Madasu et al., 2025; Burda-Lassen et al., 2025; Ananthram et al., 2025; Yadav et al., 2025).

### 3 Methodology

Following prior work on activation-based neuron analysis (Huo et al., 2024; Huang et al., 2024; Fang et al., 2024; Tang et al., 2024), we use a three-stage pipeline (Figure 2): (1) record decoder-MLP neuron activations on culturally grounded VQA items that the unablated model answers correctly;

(2) score and select culture-sensitive neurons using the identification methods in §3.2; and (3) causally test these neurons by inference-time deactivation and measure culture-specific performance changes.

#### 3.1 Step 1: Recording Activations

We instrument the decoder MLPs of each VLM and record neuron activations on VQAs that the unmasked model answers correctly. The assumption is that neurons that are preferentially active when processing information tied to a particular culture will display distinctive activation patterns on the respective culture’s inputs.

**SwiGLU activations.** We focus on the decoder MLP nonlinearity branch in the SwiGLU block (Shazeer, 2020). Let  $h_{l-1,t} \in \mathbb{R}^d$  be the layer- $l-1$  hidden state at token position  $t$ . A standard SwiGLU MLP computes

$$u_{l,t} = h_{l-1,t}W_u + b_u, \quad v_{l,t} = h_{l-1,t}W_v + b_v, \quad (1)$$

$$g_{l,t} = \text{SiLU}(u_{l,t}), \quad z_{l,t} = (g_{l,t} \odot v_{l,t})W_o + b_o, \quad (2)$$

where  $\text{SiLU}(x) = x \sigma(x)$  and  $\sigma(\cdot)$  is the sigmoid. For each neuron (dimension)  $n$  in layer  $l$  and token  $t$ , we denote the recorded scalar activation by

$$a_{l,n,t} = (g_{l,t})_n. \quad (3)$$

Because  $\sigma(x) > 0$  for all  $x$ ,  $\text{SiLU}(x)$  has the same sign as  $x$ , so  $\mathbb{I}(a_{l,n,t} > 0)$  is equivalent to  $\mathbb{I}((u_{l,t})_n > 0)$ .

**Valid-token masking.** Let  $m_{i,t} \in \{0, 1\}$  be a valid-token mask for example  $i$  at token position  $t$ , consistent with each model’s internal attention mask. This excludes padding and special markers (e.g., image delimiters), while retaining both text tokens and visual tokens consumed by the decoder. We instrument only the decoder (not upstream vision encoders).

**Activation statistics.** Let  $\mathcal{C}$  be the set of cultures and let  $\mathcal{I}_c$  be the set of correctly answered identification examples tagged with culture  $c$ . We accumulate per-neuron, per-culture sufficient statistics:

$$K_{l,n}^{(c)} = \sum_{i \in \mathcal{I}_c} \sum_t m_{i,t} \mathbb{I}(a_{l,n,t}^{(i)} > 0), \quad (4)$$

$$T_c = \sum_{i \in \mathcal{I}_c} \sum_t m_{i,t}. \quad (5)$$

Here  $K$  counts how often a neuron fires positively, and  $T_c$  is the total number of valid tokens observed for culture  $c$ . Restricting to correctly answered examples reduces noise from activations associated with failures.

### 3.2 Step 2: Identification of Culture-Sensitive Neurons

Using the aggregated statistics from Step 1, we define the token-level activation probability:

$$P_{l,n}^{(c)} = \frac{K_{l,n}^{(c)}}{T_c}. \quad (6)$$

#### 3.2.1 Baseline Identification Methods

As our identification methods for neuron scoring, we consider the following existing baseline methods. For each culture  $c$ , these methods return a ranking of neuron indices from most to least selective. Deciding how many of those neurons to select as culture-sensitive is a hyperparameter setting. To allow for a fair comparison across methods, we select the  $r\%$  highest scoring neurons out of all MLP-neurons as culture-sensitive, where we set  $r=1$  throughout.

**Random Selection (RND)** To evaluate if cultural subsets are inherently sensitive to arbitrary masking, we use a global random baseline that samples a fixed total number of neurons (i.e.,  $r\%$ ) uniformly from all layers, independent of culture. This produces a single mask shared across cultures and is compute-efficient, without enforcing any layer-wise quota.

**Activation Probability (LAP)** (Gurnee et al., 2024; Voita et al., 2024). LAP selects neurons that frequently fire for a given target (here adapted to culture), which emphasizes firing frequency alone. Directly using the activation probability:

$$s_{l,n}^{\text{LAP}}(c) = P_{l,n}^{(c)}. \quad (7)$$

For each  $c$ , we rank neurons by  $s_{l,n}^{\text{LAP}}(c)$  and select the top  $r\%$ .

**Activation Probability Entropy (LAPE)** (Tang et al., 2024; Huo et al., 2024; Namazifard and Poech, 2025). LAPE measures how selectively a neuron fires across cultures by computing the entropy of its culture profile. For each neuron, define the (approximately) normalized culture profile:

$$\tilde{P}_{l,n}^{(c)} = \frac{P_{l,n}^{(c)}}{\sum_{c' \in \mathcal{C}} P_{l,n}^{(c')} + \epsilon}, \quad \epsilon > 0, \quad (8)$$

and Shannon entropy:

$$s_{l,n}^{\text{LAPE}} = - \sum_{c \in \mathcal{C}} \tilde{P}_{l,n}^{(c)} \log \tilde{P}_{l,n}^{(c)}. \quad (9)$$

Because  $\epsilon$  is used only for numerical stability,  $\tilde{P}_{l,n}$  is nearly normalized for active neurons (where  $\sum_{c'} P_{l,n}^{(c')} \gg \epsilon$ ). Lower entropy indicates stronger culture selectivity.

We implement the following procedure: (1) **Activity filter.** We keep neurons whose maximal firing probability exceeds a threshold:  $\max_c P_{l,n}^{(c)} > p_{\text{th}}$ , where  $p_{\text{th}}$  is set to the  $\alpha$ -percentile of all values in  $\{P_{l,n}^{(c)}\}$  (we use  $\alpha = 95$ ). Neurons failing this criterion are treated as inactive and excluded from candidate selection. (2) **Low-entropy candidate pool.** From the remaining neurons, we select a candidate pool consisting of the lowest- $\rho$  fraction by  $s_{l,n}^{\text{LAPE}}$ . In our implementation, we set  $\rho = \min(1, 5r)$ , i.e., the candidate pool size is at most five times the final selection rate. (3) **Per-culture selection.** For each culture  $c$ , we select the top- $r\%$  neurons within the candidate pool by their firing probabilities  $P_{l,n}^{(c)}$  to form the culture-specific mask. (A neuron may be selected for multiple cultures because masks are constructed independently per culture.)

**MAD (Mean Activation Difference)** (Bau et al., 2019; Dalvi et al., 2019). We use the same post-nonlinearity activations  $a_{l,n,t}^{(i)}$  and valid-token mask

$m_{i,t}$  as in Step 1. Define the across-culture mean firing probability:

$$\bar{P}_{l,n} = \frac{1}{|C|} \sum_{c \in C} P_{l,n}^{(c)}. \quad (10)$$

The MAD score for culture  $c$  is the absolute deviation from this mean:

$$s_{l,n}^{\text{MAD}}(c) = \left| P_{l,n}^{(c)} - \bar{P}_{l,n} \right|. \quad (11)$$

For each culture  $c$ , we rank neurons by  $s_{l,n}^{\text{MAD}}(c)$  and select the top  $r\%$ .

### 3.2.2 Contrastive Activation Margin (ConAct)

In a preliminary analysis, we compute each neuron’s standard deviation of  $P_{l,n}^{(c)}$  across cultures and observe that in Qwen2.5-VL-7B and Pangea-7B, a substantial fraction of neurons (12.27% and 9.57%) satisfy  $\text{std}_c(P_{l,n}^{(c)}) > \text{mean}_c(P_{l,n}^{(c)})$  (Appendix D Table 6), exhibiting high activation variance across cultures. This suggests that a large mean-based difference may not necessarily indicate cultural specialization but may arise from high intrinsic variability.

To mitigate this, we introduce *Contrastive Activation Margin* (ConAct), a margin-based selector that measures the gap between the most-activated culture and its nearest competitor. By focusing on this contrast rather than deviation from the mean, ConAct is less sensitive to global variance and is expected to be more effective in high-variance models. We thus hypothesize that deactivating ConAct-identified neurons will lead to a larger culture-specific performance drop in such models, while in low-variance models, ConAct and MAD will likely identify similar neurons. For each neuron  $(l, n)$ , define the top culture and runner-up by activation probability:

$$c_{l,n}^{(1)} = \arg \max_{c \in C} P_{l,n}^{(c)}, \quad P_{l,n}^{(1)} = \max_{c \in C} P_{l,n}^{(c)}, \quad (12)$$

$$P_{l,n}^{(2)} = \max_{c \in C \setminus \{c_{l,n}^{(1)}\}} P_{l,n}^{(c)}. \quad (13)$$

The ConAct score assigns each neuron exclusively to its top culture:

$$s_{l,n}^{\text{ConAct}}(c) = \begin{cases} P_{l,n}^{(1)} - P_{l,n}^{(2)}, & \text{if } c = c_{l,n}^{(1)}, \\ -\infty, & \text{otherwise.} \end{cases} \quad (14)$$

If we use an aggregated culture group (e.g., pooling multiple country–language pairs), we treat it as a single culture in  $C$  by pooling its member pairs at the data level (i.e.,  $I_c$  is the union of examples in that group). All sufficient statistics are then computed on the pooled set. ConAct is applied unchanged on the resulting  $\{P_{l,n}^{(c)}\}_{c \in C}$ .

### 3.3 Step 3: Intervention through Neuron Deactivation

We causally test whether neurons selected in Step 2 are important for culture-sensitive behavior by deactivating them at inference time and measuring the impact on the evaluation split.

Let  $\mathcal{M}_l^{(m, c_{\text{src}})} \subseteq \{1, \dots, D_l\}$  be the set of neuron indices selected by method  $m$  for source culture  $c_{\text{src}}$  at decoder layer  $l$  (where  $D_l$  is the SwiGLU hidden width). We form a binary keep-mask  $r_l^{(m, c_{\text{src}})} \in \{0, 1\}^{D_l}$ :

$$r_{l,n}^{(m, c_{\text{src}})} = \begin{cases} 0, & n \in \mathcal{M}_l^{(m, c_{\text{src}})} \quad (\text{deactivate}) \\ 1, & \text{otherwise.} \end{cases} \quad (15)$$

During inference, we apply this mask to the SwiGLU nonlinearity output (broadcast over tokens):

$$\tilde{g}_{l,t} = g_{l,t} \odot r_l^{(m, c_{\text{src}})}, \quad (16)$$

and replace Eq. (2) with

$$z_{l,t} = (\tilde{g}_{l,t} \odot v_{l,t}) W_o + b_o, \quad (17)$$

leaving all other components unchanged. We then compare masked vs. unmasked generations on each evaluation culture to quantify culture-specific effects.

## 4 Experimental Setup

### 4.1 Dataset and Culture Grouping

We employ the CVQA dataset (Romero et al., 2024) as our testbed and operationalize “culture” through CVQA’s country–language pair (e.g., “Ireland–Irish”) taxonomy. Each item is a VQA question paired with an image and tagged by a country–language pair. We treat most country–language pairs as standalone cultures. Additionally, to study grouping effects, we form three aggregated culture groups by pooling pairs that share a country tag (India-all; Indonesia-all) or a language tag (all-Spanish). A subset of the culture groups and their question counts is shown in Table 1 (full list in Appendix A.2).

CVQA Pairs	Cultures	# Qs (I)	# Qs (E)
Brazil–Portuguese	BRA	142	142
Bulgaria–Bulgarian	BGR	185	186
China–Chinese	CHN	155	156
Egypt–Egyptian Arabic	EGY	101	102
Ethiopia–Amharic	ETA	117	117
Ethiopia–Oromo	ETO	107	107
France–Breton	FRA	202	203
India–Bengali		143	143
India–Hindi		100	101
India–Marathi	IND	101	101
India–Tamil		107	107
India–Telugu		100	100
India–Urdu		110	110
Indonesia–Indonesian		206	206
Indonesia–Javanese	IDN	148	149
Indonesia–Minangkabau		125	126
Indonesia–Sundanese		100	100
Ireland–Irish	IRL	163	163
...	...	...	...
Total		5178	5196

Table 1: **Culture subset and VQA statistics.** CVQA country–language pairs with [“India”, “Indonesia”] country tag are assigned to one of the grouped cultures. # Qs (I) denotes the number of questions used for activation recording and neuron identification, while # Qs (E) denotes the number of questions used for masked generation. Full table in Appendix A.2.

To minimize confounding from language proficiency, we use the dataset’s prepared English translations for both questions and answer options. Moreover, this mitigates the concern of identifying language rather than culture-sensitive neurons. We split the dataset approximately 50/50 into identification and evaluation subsets, where the identification split is used exclusively for activation logging (Step 1) and the evaluation split for masked generation and evaluation (Step 3).

## 4.2 Models

We evaluate three widely used VLMs: (1) LLaVA-v1.6-Mistral-7B (Liu et al., 2023; Jiang et al., 2023), (2) Pangea-7B (Yue et al., 2025), and (3) Qwen2.5-VL-7B (Bai et al., 2025) (versions and sources can be found in Appendix A Table 3).

The selected models differ in backbone, supervision, and cultural/linguistic coverage, allowing us to test whether culture-sensitive neurons emerge consistently across architectures and training paradigms. Moreover, we selected Pangea-7B because it was developed to support broad multi-

lingual multimodal coverage, making it a natural candidate for evaluating culture-linked behavior across diverse regions.

## 4.3 Prompting and Decoding

We use a fixed multiple-choice instruction template (Appendix A.3) for all models, requiring the output to be the complete option content rather than the label. Maximum generation length is set to 20, which is sufficient to return a full answer-option span. Decoding is deterministic (temperature 0; no sampling). Generations violating the format are normalized by the extraction heuristic (Appendix A.4).

## 4.4 Measuring Cultural Sensitivity

Using each neuron selector outlined in § 3, we obtain a set of culture-sensitive neurons for each source culture  $c_{\text{src}} \in \mathcal{C}$ . To evaluate to what extent these neurons are indeed culture-sensitive, we study two conditions: (1) **Self-deactivation:**  $c_{\text{src}} = c_{\text{eval}}$ , where the same culture from which the neurons were identified was used for evaluation. (2) **Cross-deactivation:**  $c_{\text{src}} \neq c_{\text{eval}}$ , where the neurons were identified from a culture that differs from the one under evaluation. This design allows us to test whether the selected neurons are primarily associated with a particular culture rather than affecting the model’s overall capacity.

**Metrics.** We assess each condition using two complementary metrics: (1) **Accuracy change** ( $\Delta$ ) measures the change in task performance between the full model and the masked model on a particular subset. Let  $Acc_{\text{full}}$  and  $Acc_{\text{masked}}$  denote accuracies reported in percentage (i.e., in  $[0, 100]$ ). We define the accuracy change as the percentage-point difference. (2) **Flip rate** is the proportion of items whose predicted answers differ from the full model, revealing decision shifts regardless of accuracy. Concretely, let  $\hat{a}_i^{\text{mask}}$  and  $\hat{a}_i^{\text{full}}$  denote the model’s predicted answers with and without ablation masking (after normalization; Appendix A.4) for item  $i$ , respectively.

$$\Delta \text{Acc} = \text{Acc}_{\text{masked}} - \text{Acc}_{\text{full}},$$

$$\text{FlipRate} = 100 \cdot \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{a}_i^{\text{full}} \neq \hat{a}_i^{\text{mask}}].$$

**Interpretation.** Ablating culture-sensitive neurons should harm performance when the evaluation culture matches the source culture (large negative  $\Delta$ , high flip rate), but have minimal effect otherwise ( $\Delta$  and flip rate close to 0). Hence, we focus

VLM	Metric	Eval. Setting	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-VL-7B	Acc. $\Delta$	Self-Deactivation	-0.19	+0.96	+0.56	-4.64	<b>-5.52</b>
		Cross-Deactivation Avg.	-	+1.07	<b>+0.61</b>	-1.31	-0.64
		Self-Cross Gap	-	-0.08	-0.05	-3.33	<b>-4.88</b>
	Flip Rate	Self-Deactivation	4.66	<b>17.05</b>	4.64	12.03	12.61
		Cross-Deactivation Avg.	-	17.21	<b>4.12</b>	5.96	4.25
		Self-Cross Gap	-	-0.16	+0.52	+6.07	<b>+8.36</b>
Pangea-7B	Acc. $\Delta$	Self-Deactivation	1.02	+1.00	-0.74	-4.20	<b>-4.33</b>
		Cross-Deactivation Avg.	-	+0.89	<b>-0.37</b>	-1.34	-0.72
		Self-Cross Gap	-	+0.11	-0.38	-2.86	<b>-3.61</b>
	Flip Rate	Self-Deactivation	6.45	<b>24.10</b>	6.52	13.55	12.99
		Cross-Deactivation Avg.	-	23.82	<b>6.18</b>	8.80	7.34
		Self-Cross Gap	-	+0.28	+0.34	+4.75	<b>+5.65</b>
LLaVA-v1.6-Mistral-7B	Acc. $\Delta$	Self-Deactivation	-0.50	-2.50	<b>-4.43</b>	-1.46	-1.39
		Cross-Deactivation Avg.	-	-2.74	-4.44	<b>-0.53</b>	-0.63
		Self-Cross Gap	-	+0.24	+0.01	<b>-0.93</b>	-0.76
	Flip Rate	Self-Deactivation	7.01	<b>17.82</b>	11.44	7.74	9.58
		Cross-Deactivation Avg.	-	17.74	11.28	<b>6.56</b>	7.63
		Self-Cross Gap	-	+0.08	+0.15	+1.18	<b>+1.95</b>

Table 2: **Ablation results on CVQA using culture-sensitive neurons selected by five identification methods.**

We report signed changes of accuracy change and flip rate relative to the unablated full model. Self-deactivation is measured on the culture used for identification (diagonal), and cross-deactivation on other cultures (off-diagonal). RND (Random Selection) is not a culture-specific masking and hence does not distinguish “self-” or “cross-” results. We use the self-cross gap to summarize cultural specificity; larger negative gap indicates stronger culture-specific impact with less spillover. We bold the method with the smallest cross effect, largest self-effect and self-cross gap.

on the gap between the self-deactivation effect and the average cross-deactivation effect as the main indicator of cultural sensitivity. Methods that yield larger self-cross gaps better isolate neurons that are critical, yet relatively specific to a given culture. Larger gaps (more negative) indicate stronger culture-specific impact with less spillover.

## 5 Results

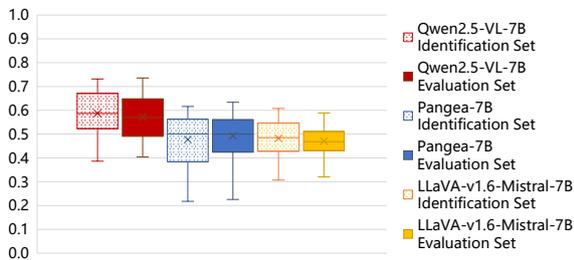


Figure 3: **Unablated full models per-culture accuracy on CVQA.** Distribution of per-culture accuracies for the three models on the identification split (marked in dots) and the evaluation split (marked in solid color). The full table of per-culture results appears in Appendix D.

### 5.1 Baseline Model Performance on CVQA

We first assess the unablated full model performance on CVQA, shown in Figure 3. All three VLMs exhibit substantial variation in performance across cultures. Qwen2.5-VL-7B achieves the highest median accuracy ( $\approx 0.60$ ), while Pangea-7B and LLaVA-v1.6-Mistral-7B reach around 0.50. Importantly, identification and evaluation splits yield similar performance, suggesting that subsequent ablation results are not confounded by train-test mismatch. Overall, the models show uneven cultural competence but stable baselines, providing a reliable reference point for neuron ablations.

### 5.2 Culture-Sensitive Neuron Ablation

Table 2 reports accuracy change ( $\Delta$ ) and flip rates when deactivating neurons selected by each identification method (§3.2). We analyze two evaluation settings as defined in § 4.4: *self-deactivation* (masking neurons identified for the same culture as the evaluation set) and *cross-deactivation* (masking neurons identified for a different culture).

For Qwen2.5-VL-7B and Pangea-7B, ConAct yields the largest self-deactivation drops in accuracy (Qwen:  $-5.52$ ; Pangea:  $-4.33$ ) paired with small cross-deactivation changes ( $< 1$ ), showing

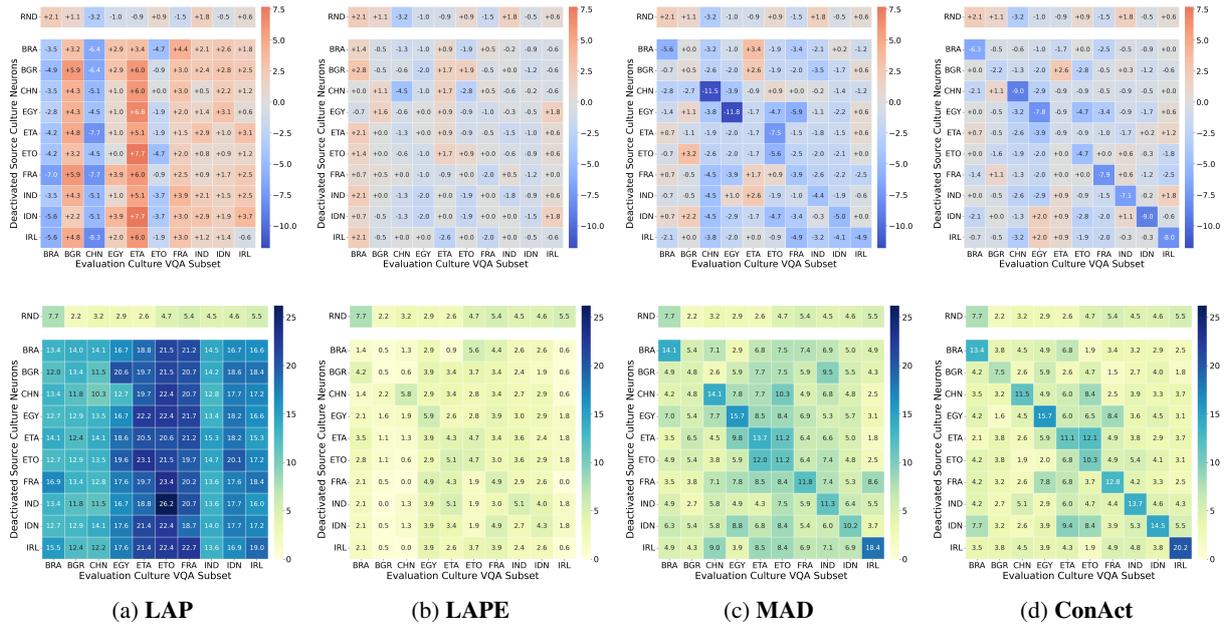


Figure 4: **Accuracy change**  $\Delta$  (top) and **flip rate** heatmaps (bottom) on CVQA for different identification methods (Qwen2.5-VL-7B; showing the first ten evaluated cultures). On the y-axis we have the source culture for which neurons are identified and ablated and on the x-axis the culture used for evaluation. We report signed changes relative to the unablated full model. Diagonal cells show self-deactivation results.

that the selected neurons are both important and relatively specific to their source culture. The associated flip rate gaps are likewise large and positive, indicating that predictions change substantially only within the target culture. By contrast, LAP and MAD often produce broader off-diagonal interference, capturing neurons linked to shared or generic multimodal cues rather than culture-specific signals. Occasionally, LAP even improves performance upon masking. Such gains can arise in ablation studies when removed units encode spurious or overly dominant features, effectively acting as a form of pruning (Ali et al., 2025). For LLaVA-v1.6-Mistral-7B, LAPE induces the strongest self-deactivation drop ( $-4.43$ ) but also larger cross-cultural spillover. ConAct and MAD yield neurons that show more cultural selectivity yet smaller in effect magnitude.

**Culture-Specific patterns.** Figure 4 visualizes accuracy changes ( $\Delta$ ) and flip rates for Qwen2.5-VL-7B of the first ten cultures. We find that LAP shows broad column-shaped reductions (large off-diagonals), pointing to less specific features; LAPE reveals fairly little selectivity. On the contrary, both MAD and ConAct produce sharp diagonal degradations with limited but non-negligible off-diagonal changes, while ConAct achieves the cleanest separation between self and cross conditions.

The results for MAD and ConAct evidence strong mapping between masked neuron sets and cultures. Additionally, some geographically linked cultures show correlated effects. For instance, deactivating EGY-neurons yields significant impact on ETO, another African culture group. Interestingly, for some cultures we observe small accuracy gains even when ablating neuron sets identified from other cultures (e.g., BGR, ETA). A plausible explanation is that certain selected units capture spurious cues that hurt generalization on those subsets; removing them can therefore resemble pruning rather than culture-targeted disruption (Ali et al., 2025).

### 5.3 Distribution Patterns across Layers

Figure 5 shows the layer-wise distribution of culture-sensitive neurons identified in Qwen2.5-VL-7B (28-layer decoder). Understanding where such neurons concentrate within the network can offer clues about how culture-related information is integrated e.g., whether it is handled early, during basic feature fusion, or later, during high-level reasoning. Moreover, comparing distributions across identification methods reveals whether different methods capture similar or distinct functional subspaces, while cross-cultural differences can hint at culture-specific processing pathways.

We observe that culture-sensitive neurons generally cluster in the first layer (layer 0) and the

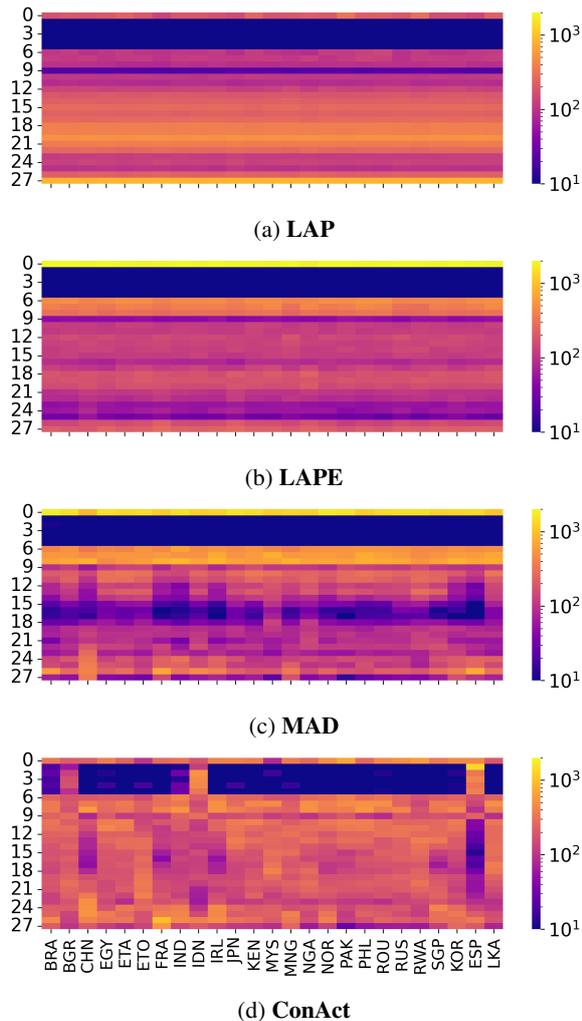


Figure 5: **Layer-wise counts of identified neurons** by different methods (Qwen2.5-VL-7B; log-scaled color).

early-mid layers (6–8), with relatively sparse presence in deeper blocks. Interestingly, MAD tends to bypass the central layers (15–18), whereas ConAct identifies neurons more evenly across mid-to-late layers. ConAct also shows culture-specific deviations, for example, in BGR and IDN, layers 6–8 contain a higher proportion of selected neurons than in other cultures. These patterns suggest that both the choice of method and culture influence which layers of the model are most engaged in culturally grounded processing.

#### 5.4 Effect of Ablation on Model Behavior

Figure 1 highlights two key observations about how ablation disrupts model behavior. First, we do not observe widespread degradation of format compliance under decoder-level masking: masked generations typically remain compatible with the multiple-choice response format (Appendix A.4),

suggesting that the intervention does not generally collapse task framing at the sparsity levels we study. Second, we find that different identification methods perturb cultural knowledge in distinct ways: RND yields only small changes, suggesting that arbitrary neurons are rarely detrimental for culture-specific performance. In contrast, LAPE and ConAct push the model to different incorrect but plausible options. This suggests that the ablated neurons induce selective culture degradation.

Overall, our analyses reveal several consistent patterns across 25 cultural groups and three model architectures: (1) A select subset of decoder neurons exhibit clear culture-sensitive activation patterns, suggesting that cultural knowledge is at least in part encoded locally. (2) These neurons play an important role in culturally grounded processing: their removal selectively degrades performance on the corresponding culture while largely preserving performance elsewhere. (3) Culture-sensitive neurons are not uniformly distributed but cluster in early to mid decoder layers (with some model- and culture-specific variation). (4) Among all identification methods, ConAct most effectively isolates such neurons.

## 6 Conclusion

This study provides empirical evidence for the existence of culture-sensitive neurons in VLMs by showing inference-time ablations of targeted subsets of neurons that selectively disrupt VLMs’ culture-specific performance. We introduce a margin-based selector (ConAct) that allows for more precise identification of culture-sensitive neurons. Among the identification methods we compare, ConAct identifies neurons whose ablation yields the largest self-deactivation drops with minimal cross-deactivation spillover on Qwen2.5-VL-7B and Pangea-7B, while LLaVA-v1.6-Mistral-7B shows resistance to specific targeting. Layer-wise analyses further show that the identified neurons concentrate in specific decoder regions, with distributions that vary by model and selector. Overall, our findings suggest that small, targeted neuron suites can serve as a diagnostic handle for probing culturally grounded behavior in VLMs, and that margin-based selection can better isolate units whose ablation yields large self-cross gaps in some architectures. Future work could extend the search beyond decoder MLPs and pair identification with activation steering.

## Limitations

**Defining “culture”.** We use CVQA’s country–language taxonomy and, for fairness to multilingual models, only the English-translated prompts to better decouple language skill from cultural recognition. This choice makes the construct closer to visual cultural knowledge than to culture-as-language-practice (Kramsch, 2014). For multilingual models, it remains unknown whether our observations would still emerge, which we leave for future work.

**Model components.** Our analysis is restricted to decoder MLP neurons and does not cover attention heads, vision encoders, or alignment modules, which may also encode culture-sensitive behavior. We rely on activation-frequency summaries rather than more fine-grained temporal or token-level dynamics, and we fix hyperparameters for neuron selection based on computational budget.

## Ethical Considerations

This study aims to improve transparency and fairness in multimodal models by examining culture-sensitive neurons. All experiments are conducted on a publicly available dataset (CVQA), and no new human subject data or personally identifiable information is used.

A potential ethical concern lies in the definition of “culture.” For experimental feasibility, we adopt CVQA’s taxonomy of country–language pairs and, in some cases, group multiple pairs that share a common country or language tag. Such grouping is a dataset-driven simplification and does not reflect the diversity, fluidity, or internal variation within cultural communities. Our results should not be interpreted as essentializing or stereotyping real-world cultures but rather as insights into how models respond to the categories provided by the benchmark.

The methods presented are intended for diagnostic use only. While they can help reveal and quantify cultural disparities in model behavior, they are not in themselves fairness interventions. Misuse of these methods to draw normative claims about communities would be harmful and contrary to the goals of this work. We encourage future studies to incorporate broader and more inclusive datasets when assessing and mitigating cultural bias in multimodal systems.

## Acknowledgments

We thank Simon King, Korin Richmond, and Catherine Lai at the University of Edinburgh for their constant support during the course of the project. Special thanks to Jinzuomu Zhong for providing help on computational resources.

## References

- Ameen Ali Ali, Shahar Katz, Lior Wolf, and Ivan Titov. 2025. [Detecting and pruning prominent but detrimental neurons in large language models](#). In *Second Conference on Language Modeling*.
- Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. [See it from my perspective: How language affects cultural bias in image understanding](#). In *The Thirteenth International Conference on Learning Representations*.
- Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. [Composable sparse fine-tuning for cross-lingual transfer](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. [Identifying and controlling important neurons in neural machine translation](#). In *International Conference on Learning Representations*.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. [Network Dissection: Quantifying Interpretability of Deep Visual Representations](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, Los Alamitos, CA, USA. IEEE Computer Society.
- David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. [Understanding the role of individual units in a deep neural network](#). *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.
- Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. [BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

- Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How culturally aware are vision-language models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 1–6. IEEE.
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. [What is one grain of sand in the desert? analyzing individual neurons in deep nlp models](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.
- Junfeng Fang, Zongze Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multimodal large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. [Universal neurons in GPT2 language models](#). *Transactions on Machine Learning Research*.
- Geert Hofstede. 1980. Culture and organizations. *International studies of management & organization*, 10(4):15–41.
- Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024. [Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models](#). *Preprint*, arXiv:2410.04819.
- Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. [MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.
- Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. [Style-specific neurons for steering LLMs in text style transfer](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [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.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Avinash Madasu, Vasudev Lal, and Phillip Howard. 2025. [Cultural awareness in vision-language models: A cross-country exploration](#). In *CVPR 2025 Workshop Vision Language Models For All*.
- Danial Namazifard and Lukas Galke Poch. 2025. [Isolating culture neurons in multilingual large language models](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 768–785, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. [Towards interpreting visual information processing in vision-language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. [Grounding multilingual multimodal LLMs with cultural knowledge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.
- Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. [Finding and editing multi-modal neurons in pre-trained transformers](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037, Bangkok, Thailand. Association for Computational Linguistics.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrana, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. [Survey of cultural awareness in language models: Text and beyond](#). *Computational Linguistics*, 51(3):907–1004.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan

- Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. [Cvqa: Culturally-diverse multilingual visual question answering benchmark](#). *Preprint*, arXiv:2406.05967.
- Noam Shazeer. 2020. [Glu variants improve transformer](#). *Preprint*, arXiv:2002.05202.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Edward Burnett Tylor. 1871. *Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom*, volume 2. J. Murray.
- Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. [Neurons in large language models: Dead, n-gram, positional](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.
- Jiaqi Xu, Cuiling Lan, and Yan Lu. 2025. Deciphering functions of neurons in vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3173–3181.
- Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. [Beyond words: Exploring cultural value sensitivity in multimodal models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7592–7608, Albuquerque, New Mexico. Association for Computational Linguistics.
- Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. [Neuron-level analysis of cultural understanding in large language models](#). *Preprint*, arXiv:2510.08284.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.
- Zeping Yu and Sophia Ananiadou. 2024. [Neuron-level knowledge attribution in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.
- Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. [Pangea: A fully open multilingual multimodal LLM for 39 languages](#). In *The Thirteenth International Conference on Learning Representations*.
- Xiutian Zhao, Björn Schuller, and Berrak Sisman. 2026. [Discovering and causally validating emotion-sensitive neurons in large audio-language models](#). *Preprint*, arXiv:2601.03115.
- Xiutian Zhao, Ke Wang, and Wei Peng. 2024. [Measuring the inconsistency of large language models in preferential ranking](#). In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 171–176, Bangkok, Thailand. Association for Computational Linguistics.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

## A Reproducibility

### A.1 Models and Sources

Models	Sources
LLaVA-v1.6-Mistral-7B	<a href="https://huggingface.co/llava-hf/LLaVA-v1.6-Mistral-7B-hf">https://huggingface.co/llava-hf/LLaVA-v1.6-Mistral-7B-hf</a>
Pangea-7B	<a href="https://huggingface.co/neulab/Pangea-7B">https://huggingface.co/neulab/Pangea-7B</a>
Qwen2.5-VL-7B	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct</a>

Table 3: Sources of the evaluated models.

### A.2 Culture Grouping of CVQA

The CVQA benchmark comprises 39 country-language pairs, several of which share the same country or language tags. To study potential grouping effects, we construct three aggregated culture sets that pool pairs with a shared attribute: India-all (IND) (all pairs tagged with country “India”),

Indonesia—all (IDN) (all pairs tagged with country “Indonesia”), and all–Spanish (ESP) (all pairs whose language is Spanish). Table 4 reports the mapping from individual pairs to each aggregate and the number of questions per subset in the identification and evaluation splits.

CVQA Pairs	Grouped Cultures	# Qs (I)	# Qs (E)
Brazil–Portuguese	BRA	142	142
Bulgaria–Bulgarian	BGR	185	186
China–Chinese	CHN	155	156
Egypt–Egyptian Arabic	EGY	101	102
Ethiopia–Amharic	ETA	117	117
Ethiopia–Oromo	ETO	107	107
France–Breton	FRA	202	203
India–Bengali		143	143
India–Hindi		100	101
India–Marathi	IND	101	101
India–Tamil		107	107
India–Telugu		100	100
India–Urdu		110	110
Indonesia–Indonesian		206	206
Indonesia–Javanese	IDN	148	149
Indonesia–Minangkabau		125	126
Indonesia–Sundanese		100	100
Ireland–Irish	IRL	163	163
Japan–Japanese	JPN	101	102
Kenya–Swahili	KEN	136	137
Malaysia–Malay	MYS	157	158
Mongolia–Mongolian	MNG	156	156
Nigeria–Igbo	NGA	100	100
Norway–Norwegian	NOR	146	150
Pakistan–Urdu	PAK	108	108
Philippines–Filipino	PHL	101	102
Romania–Romanian	ROU	151	151
Russia–Russian	RUS	100	100
Rwanda–Kinyarwanda	RWA	117	118
Singapore–Chinese	SGP	106	106
South Korea–Korean	KOR	145	145
Argentina–Spanish		132	133
Chile–Spanish		117	117
Colombia–Spanish		120	121
Ecuador–Spanish	ESP	181	181
Mexico–Spanish		161	162
Spain–Spanish		159	159
Uruguay–Spanish		157	158
Sri Lanka–Sinhala	LKA	112	113
Total		5178	5196

Table 4: **Culture subsets and VQA statistics.** CVQA country–language pairs with “Spanish” language tag or [“India”, “Indonesia”] country tag are assigned to one of the aggregated cultures, and other pairs remain stand-alone. # Qs (I) denotes the number of questions used for activation recording and neuron identification, while # Qs (E) denotes the number of questions used for masked generation and evaluation.

### A.3 Prompt Template for Multiple-Choice VQA

Listing 1: Prompt template used for VQA generation

Answer the following multiple-choice question based on the image.

Question:  
{question}

Options:  
{option 1}  
{option 2}  
{option 3}  
{option 4}

Your response must be ONLY the text of the correct option from the list above, and nothing else.

### A.4 Answer Normalization Process

To ensure reliable evaluation of model predictions in the multiple-choice setting, we implemented a normalization procedure to mitigate inconsistencies in the format and phrasing of generated outputs.

First, the prediction string is converted to lowercase and standardized by collapsing all whitespace into single spaces and trimming leading and trailing spaces. This step minimizes mismatches caused by case sensitivity or formatting irregularities. Each answer option is similarly normalized to lowercase. The algorithm then searches for whole-word matches of each choice within the normalized prediction using word-boundary matching to prevent false positives. Because LLMs are known to exhibit label bias in multiple-choice answering settings (Zheng et al., 2024; Zhao et al., 2024), we require the model to output the full content of the chosen option rather than its label (e.g., “A”, “B”).

Although the prompt explicitly instructs the model to generate a single answer (Appendix A.3), instruction-tuned language models may still produce extended reasoning, which makes simple substring matching insufficient. Therefore, we applied a heuristic when multiple choices appear in the output: the last-mentioned choice is treated as the model’s final decision. This heuristic reflects the common generation pattern where models deliberate over several options before declaring a final answer (e.g., “Option A is plausible, but B is incorrect, so the answer is C”). If multiple choices appear in the output, we treat the last-mentioned choice as the model’s final decision, reflecting common generations that deliberate over several op-

tions before committing to an answer. If no choice can be confidently identified using whole-word matching and the last-mentioned rule above, we count the prediction as incorrect for accuracy.

This two-stage normalization and extraction process improves the evaluation’s robustness to varied model output styles while prioritizing the most plausible interpretation of the model’s intended final answer.

## B Layer-Wise Neuron Distribution

We report the layer-wise counts of culture-sensitive neurons identified in LLaVA-v1.6-Mistral-7B (Figure 6) and Pangea-7B (Figure 7), aggregated over all selected cultures. To improve visual comparability across layers, all heatmaps use a logarithmic color scale, which compresses extremely large counts in dominant layers while expanding the dynamic range for smaller counts elsewhere.

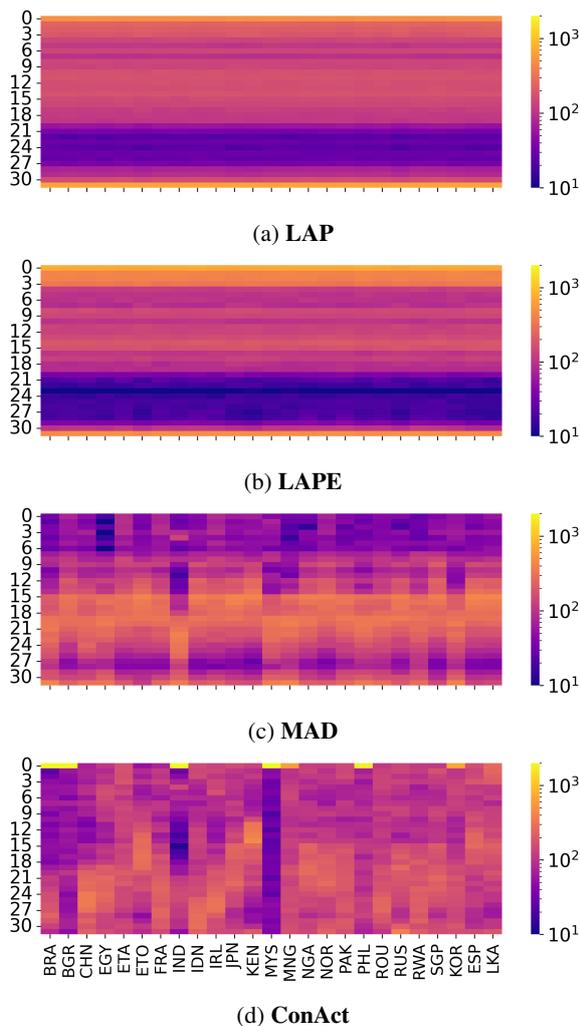


Figure 6: Layer-wise counts of identified neurons (LLaVA-v1.6-Mistral-7B).

**LLaVA-v1.6-Mistral-7B** Culture-sensitive neuron distributions in LLaVA-v1.6-Mistral-7B vary more across identification methods than in Qwen2.5-VL-7B and Pangea-7B. LAP and LAPE concentrate strongly in the earliest and latest layers, while still covering a broad early-to-mid region (1–20). In contrast, both show a pronounced low-activation “band” around 21–27, indicating limited selection in those layers. MAD primarily concentrates in mid-to-late layers (15–24). ConAct is generally sparser than MAD and selects relatively more neurons in early layers (0–6), suggesting a different locus of culture-selective evidence under the margin-based criterion.

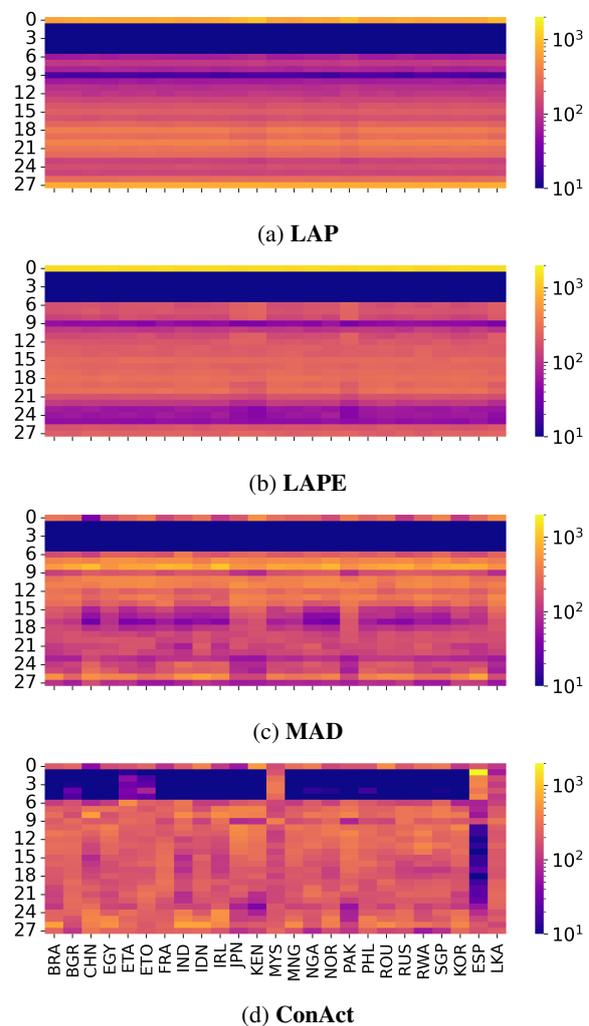


Figure 7: Layer-wise counts of identified neurons (Pangea-7B).

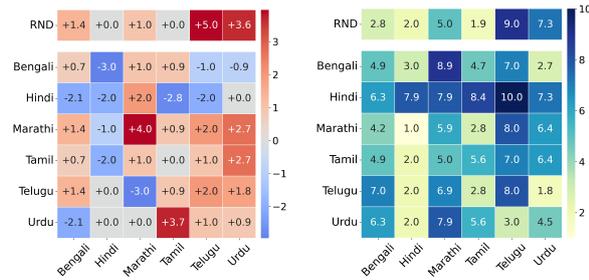
**Pangea-7B** The layer-wise distribution for Pangea-7B closely mirrors Qwen2.5-VL-7B, consistent with the fact that Pangea-7B’s language component is built on a Qwen2-7B-Instruct backbone (Yang et al., 2024; Yue et al., 2025), a di-

rect predecessor of Qwen2.5-7B-Instruct (Qwen et al., 2025) used by Qwen2.5-VL-7B. Both models share a 28-layer decoder architecture. Similar to Qwen2.5-VL-7B, the aggregated all-Spanish culture group (ESP) exhibits a distinctive early-layer concentration (0–5) compared to other cultures, suggesting that some shared language-associated or region-associated cues are preferentially localized to early decoder MLPs in these backbones.

## C Granularity of Culture Grouping

We provide two finer-grained analyses that probe culture grouping at different granularities using Qwen2.5-VL-7B with ConAct-selected neurons: (1) within a single country across multiple languages (India), and (2) across multiple countries within a shared language (Spanish).

### C.1 Within a Single Country Across Languages: India



(a) Accuracy changes  $\Delta$  for India cultures. (b) Flip rates for India cultures.

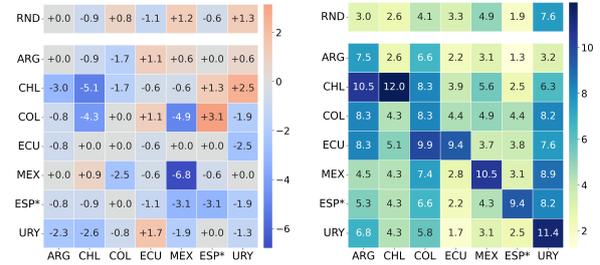
Figure 8: Within-country (India) cross-language ablations (ConAct-selected neurons on Qwen2.5-VL-7B).

Figure 8 compares ablation effects across six India language-culture subsets (Bengali, Hindi, Marathi, Tamil, Telugu, Urdu). Overall self-deactivation effects on accuracy are modest, ranging from  $-3.0$  to  $+4.0$ , but are more consistently negative for Hindi ( $-2.0$ ), suggesting that the identified Hindi-specific neurons carry relatively stronger self-specific evidence.

The same-country interactions are mixed and occasionally beneficial (e.g., ablating Urdu-identified neurons yields  $+3.7$  on Tamil; ablating Marathi-identified neurons yields  $+4.0$  on Marathi), which is consistent with partial feature overlap and potential interference effects under shared national context. Such positive gains may also reflect pruning of detrimental or noisy features (Ali et al., 2025). Flip rates align with this interpretation: Hindi and Telugu exhibit higher self flip rates (7.9 and 8.0),

while off-diagonal flips are typically lower (about 1.5–6.8).

### C.2 Across Countries Within a Shared Language: Spanish



(a) Accuracy changes  $\Delta$  for Spanish-speaking cultures. (b) Flip rates for Spanish-speaking cultures.

Figure 9: Same-language (Spanish) cross-country ablations for Argentina, Chile, Colombia, Ecuador, Mexico, Spain, and Uruguay (ConAct-selected neurons on Qwen2.5-VL-7B). “ESP\*” denotes Spain alone, distinct from “ESP” used elsewhere to denote the aggregated all-Spanish group.

Figure 9 reports ablation outcomes across seven Spanish-speaking national variants. Self-deactivation accuracy drops range from negligible to moderate, with the largest drop for Mexico ( $-6.8$ ), followed by Chile ( $-5.1$ ) and Uruguay ( $-2.3$ ). Flip rates indicate stronger national specificity than accuracy alone: Chile and Uruguay show high self flip rates (12.0 and 11.4), while cross-country flip rates often remain substantial (roughly 7.5–12). This pattern suggests a combination of broadly shared features (e.g., language-linked semantics) and country-distinct cues (e.g., place-specific iconography and landscapes) that can be differentially disrupted by neuron sets identified from different national subsets.

## D Complete Unablated Results

Culture	Model	LLaVA-v1.6-Mistral-7B		Pangea-7B		Qwen2.5-VL-7B	
		Iden.	Eval.	Iden.	Eval.	Iden.	Eval.
Brazil–Portuguese		0.5352	0.5493	0.5704	0.6338	0.6972	0.6901
Bulgaria–Bulgarian		0.4324	0.4301	0.5189	0.4624	0.6000	0.5000
China–Chinese		0.5355	0.4679	0.5871	0.5641	0.7161	0.7308
Egypt–Egyptian Arabic		0.4851	0.4608	0.5446	0.5294	0.6634	0.5686
Ethiopia–Amharic		0.5043	0.4957	0.4701	0.4530	0.5470	0.4274
Ethiopia–Oromo		0.4766	0.4486	0.3832	0.3551	0.4673	0.5794
France–Breton		0.3762	0.3202	0.3366	0.3202	0.4703	0.4039
India–all		0.5068	0.4773	0.6021	0.5468	0.6808	0.6239
Indonesia–all		0.4594	0.3563	0.4801	0.4819	0.5250	0.5301
Ireland–Irish		0.6074	0.5890	0.5644	0.6319	0.6196	0.6196
Japan–Japanese		0.3069	0.3627	0.2178	0.2255	0.3861	0.4216
Kenya–Swahili		0.4926	0.4599	0.4412	0.3577	0.5882	0.4818
Malaysia–Malay		0.4268	0.4304	0.5605	0.4873	0.5860	0.5380
Mongolia–Mongolian		0.4295	0.4295	0.3846	0.4167	0.4551	0.4679
Nigeria–Igbo		0.5600	0.4200	0.4700	0.4000	0.5200	0.4800
Norway–Norwegian		0.5570	0.5133	0.5034	0.5133	0.5839	0.5800
Pakistan–Urdu		0.5741	0.5648	0.3796	0.6296	0.7315	0.7037
Philippines–Filipino		0.5050	0.4804	0.5644	0.4804	0.5743	0.5882
Romania–Romanian		0.4768	0.4570	0.6159	0.5563	0.6556	0.6556
Russia–Russian		0.3900	0.5500	0.5000	0.5500	0.5600	0.6400
Rwanda–Kinyarwanda		0.4103	0.4492	0.3333	0.4322	0.4444	0.5254
Singapore–Chinese		0.5849	0.5094	0.5377	0.6038	0.7170	0.7358
South Korea–Korean		0.5793	0.5034	0.5310	0.5517	0.6207	0.5655
all–Spanish		0.4557	0.5034	0.4830	0.4995	0.5871	0.5703
Sri Lanka–Sinhala		0.3839	0.5310	0.3393	0.6195	0.6786	0.6726
Avg.		0.5345	0.5102	0.5938	0.5831	0.6522	0.6395

Table 5: **Culture-specific performance.** We report accuracy on the set of questions-answer pairs used for neuron identification and evaluation respectively.

	Qwen2.5-VL-7B	Pangea-7B	LLaVa-v1.6-Mistral-7B
Mean of neuron-wise standard deviations across cultures	0.015284	0.017388	0.020062
Std. dev. of neuron-wise standard deviations across cultures	0.012924	0.018586	0.016303
Max neuron-wise standard deviation	0.170701	0.196799	0.160319
Number of neurons where std > mean activation	<b>65101 (12.27%)</b>	<b>50772 (9.57%)</b>	<b>148 (0.03%)</b>

Table 6: **Neuron-wise standard deviations across cultures** for the three evaluated models. The preliminary experiment on Qwen2.5-VL-7B and Pangea-7B yield high variances across cultures.