

Finding Culture-Sensitive Neurons in Vision-Language Models

Xiutian Zhao^{1,3}

¹University of Edinburgh

Rochelle Choenni²

²University of Amsterdam

Rohit Saxena¹

Ivan Titov^{1,2}

³Johns 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 causal tests by deactivating the neurons flagged by different 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 minimal effects on others. Moreover, we propose a new margin-based selector – *Contrastive Activation Selection (CAS)*, and show that it outperforms existing probability- and entropy-based methods in identifying culture-sensitive neurons. Finally, our layer-wise analyses reveals that such neurons tend to cluster in certain decoder layers. Overall, our findings shed new light on the internal organization of multimodal representations.

¹

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

¹Related code and data are available at <https://github.com/xiutian/vlm-culture-neuron>.

(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 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 unexplored. 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

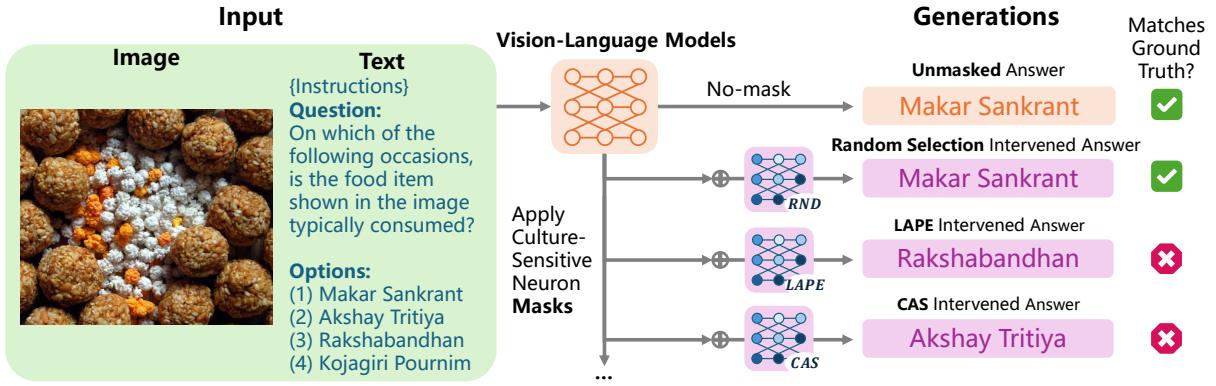


Figure 1: **An ablation example** of Qwen2.5-VL-7B on India-Marathi VQA subset. Given an image of Tilgul, an Indian sweet made from sesame seeds and jaggery, the full model selects the ground truth-matched option; **RND** mask does not affect the model’s decision, while **LAPE** and **CAS** masks redirect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

culture-selectivity, i.e. units whose activations exhibit stronger association with certain cultural contexts 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 Selection (CAS)*, 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 an well-established practice in interpreting deep neural network models. Early work on CNN 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 analyzes 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. A recent concurrent work demonstrates evidence of culture-sensitive neurons in LLMs (Namazifard and Galke, 2025). However, in vision-language multimodal settings, existing efforts are

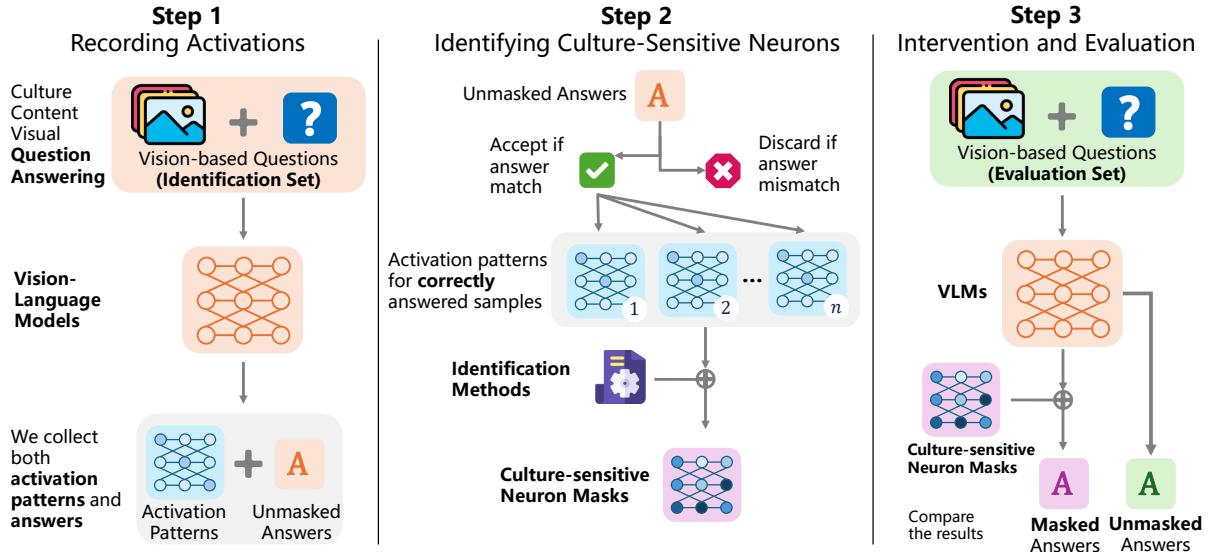


Figure 2: Pipeline for identifying and validating culture-sensitive neurons: (1) record neuron activations on culture-specific VQAs, (2) identify influential neurons using several methods, and (3) evaluate their importance by ablating the top- $r\%$ neurons and measuring the effect on accuracy and answer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

limited to identifying modality- (Huang et al., 2024; Fang et al., 2024; Xu et al., 2025) or task-specific neurons (Neo et al., 2025).

Cultural values 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 CUTLURALGROUND (de Dieu 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, 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 (Madrasu et al., 2025; Ananthram et al., 2025; Yadav et al., 2025).

3 Methodology

Following prior work on activation-based neuron analysis (Hu et al., 2024; Huang et al., 2024; Fang et al., 2024; Tang et al., 2024), we use a three-stage pipeline, illustrated in Figure 2. First, we pass culturally grounded and vision-based multiple-choice questions through each model and record neuron activation patterns as explained in § 3.1. Second, we score and select neurons for culture selectivity

using the identification methods described in § 3.2.

Finally, in § 3.3 we explain how we run intervention.

3.1 Step 1: Recording Activations

We instrument the decoder MLPs of each VLM and recording 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.

Within each decoder MLP, we monitor the non-linearity branch of the SwiGLU block (Shazeer, 2020). Concretely, given pre-activations u and v , the gated branch is $g = \text{SiLU}(u)$, which is then combined with v (Appendix B.1 Eq. 1). For each neuron n in layer l and token position t , we denote by $a_{l,n,t}$ the scalar activation corresponding to g . These values serve as the basis for all subsequent statistics.

Activation statistics. Let \mathcal{C} be the set of cultures. For a sample from culture $c \in \mathcal{C}$, and using a valid-token mask $m_t \in \{0, 1\}$ to exclude padding and special markers, we accumulate three statistics for each neuron (l, n) , note that $[x]_+ = \max(x, 0)$:

$$K_{l,n}^{(c)} += \sum_t m_t \mathbb{I}(a_{l,n,t}^{(c)} > 0),$$

$$S_{l,n}^{(c)} += \sum_t m_t [a_{l,n,t}^{(c)}]_+, \quad T_c += \sum_t m_t.$$

Here K counts how often a neuron fires positively, S measures the cumulative magnitude of its positive responses, and T_c records the total number of valid tokens for culture c . Only samples that the model answers correctly contribute, reducing noise from spurious activations. These aggregated statistics yield per-neuron activation profiles conditioned on culture, forming the foundation for our culture-sensitivity identification methods.

Text and visual tokens. In the VLMs we study, the decoder consumes a sequence that interleaves prompt tokens with visual tokens produced upstream. We instrument only the decoder, not the upstream vision encoders. Moreover, the mask m_t respects each model’s attention mask so that padding and special markers (e.g., image delimiters) are ignored.

3.2 Step 2: Identification of Culture-Sensitive Neurons

Using the counters from Step 3.1, we derive normalized statistics for each neuron–culture pair:

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

$P_{l,n}^{(c)}$ reflects how often a neuron fires on culture c ’s tokens, while $M_{l,n}^{(c)}$ couples firing frequency with activation strength.

Identification methods. As our identification methods for neuron scoring, we consider the following four existing baseline methods:

- **Random Selection (RND)** uniformly samples fixed number of neurons to evaluate if cultural subsets are inherently sensitive to arbitrary masking. This baseline is culture-independent.
- **Activation Probability (LAP, Gurnee et al., 2024; Voita et al., 2024)** ranks neurons by how often they fire for a given culture, which emphasizes firing frequency alone.

$$\text{LAP}_{l,n}^{(c)} = P_{l,n}^{(c)} = \frac{K_{l,n}^{(c)}}{T_c}$$

- **Activation Probability Entropy (LAPE, Tang et al., 2024; Namazifard and Galke, 2025)** measures how selective a neuron’s firing is across cultures by computing the entropy of its normalized activation probabilities:

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

$$\tilde{P}_{l,n}^{(c)} = \frac{P_{l,n}^{(c)}}{\sum_{c'} P_{l,n}^{(c')}}.$$

Neurons with *low* entropy (peaked distributions) are considered more culture-sensitive.

- **Mean Activation Difference (MAD, Bau et al., 2018; Dalvi et al., 2019)** incorporates magnitude as well as frequency by comparing mean positive activations for one culture against the average over the others:

$$\text{MAD}_{l,n}^{(c)} = M_{l,n}^{(c)} - \bar{M}_{l,n}^{(-c)}$$

$$\bar{M}_{l,n}^{(-c)} = \frac{1}{|\mathcal{C}| - 1} \sum_{c' \neq c} M_{l,n}^{(c')}.$$

Larger positive differences indicate neurons that fire more strongly for culture c .

Appendix B.2 provides more details on above baseline identification methods.

Neuron selection. 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 to $r=1$ in all our experiments.

3.3 Step 3: Intervention through Neuron Deactivation

We test whether the neurons selected in Step 2 are important for culture-sensitive behavior by masking them at inference time and measuring the impact on the respective culture’s evaluation subset. Let $g_{l,t} \in \mathbb{R}^{D_l}$ be the SwiGLU nonlinearity output at decoder layer l and token position t . From $\mathcal{M}_l^{(m,c_{\text{src}})}$ we form a binary *keep-mask* $r_l^{(m,c_{\text{src}})} \in \{0, 1\}^{D_l}$ where:

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

During inference we multiply the gating branch by this vector (broadcast over tokens).

3.4 Contrastive Activation Selection (CAS)

Preliminary analysis revealed that in QWEN2.5-VL-7B and PANGEA-7B, a substantial fraction of neurons (12.27% and 9.57%; Appendix E Table 6) exhibit 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 Selection (CAS)*, a margin-based selector that measures the gap between the most and the second-most active cultures for each neuron. By focusing on this contrast rather than deviation from the mean, CAS is less sensitive to global variance and is expected to be more effective in high-variance models. We thus hypothesize that deactivating CAS-identified neurons will lead to a larger culture-specific performance drop in such models, while in low-variance models, CAS and MAD will likely identify similar neurons. Using the firing probabilities $P_{l,n}^{(c)}$ from Step 1, define

$$\begin{aligned} P_{l,n}^{(1)} &= \max_{c \in \mathcal{C}} P_{l,n}^{(c)}, \quad c_{l,n}^{(1)} = \arg \max_c P_{l,n}^{(c)} \\ P_{l,n}^{(2)} &= \max_{c \in \mathcal{C} \setminus \{c_{l,n}^{(1)}\}} P_{l,n}^{(c)}, \\ s_{l,n}^{\text{CAS}}(c) &= \begin{cases} P_{l,n}^{(1)} - P_{l,n}^{(2)}, & \text{if } c = c_{l,n}^{(1)}, \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

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 keep independent pairs with unique language or country tag and aggregated pairs with shared attributes. A subset consisting of the first ten alphabetically ordered country–language pairs is reported in Table 1 (the full list is provided 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 separate the dataset by 50/50 into identification/evaluation splits: the identification split is used exclusively for activation logging (Step 1) and the evaluation split for masked generation and evaluation (Step 3).

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. Truncated for readability, full table in Appendix A.2.

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). 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 specifically to be a culturally inclusive multilingual VLM.

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 large enough (20) to return a full option token span; Decoding is deterministic (temperature 0; no sampling). generations violating the format are normalized by the extraction heuristic (Appendix A.4).

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

Table 2: **Ablation results on CVQA using culture-sensitive neurons selected by five identification methods.** As explained in § 4.4 we report two evaluation settings: Self-Deactivation and Cross-Deactivation, with Self–Cross Gaps best reflecting cultural sensitivity. The best result in each setting is boldfaced. Note that non-gap values show percentage changes relative to the unablated full model. RND (Random Selection) is not a culture-specific masking and hence does not distinguish ‘self-’ or ‘cross-’ results.

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 metrics: (1) **Accuracy change** (Δ): the difference in CVQA subset accuracy between the full model and the masked model. (2) **Flip rate**: the proportion of items whose predicted answers differ from the full model. These metrics are complementary, Δ measures the change in task performance, while flip rate reveals decision shifts even when overall accuracy remains unchanged.

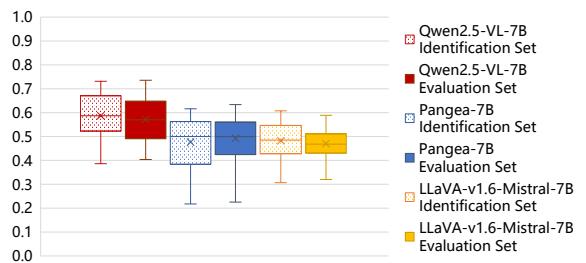


Figure 3: **Unablated full models per-culture accuracy on CVQA.** Distribution of per-culture accuracies for the three models on the identification split (marked in dots) and the evaluation split (marked in solid color). The full table of per-culture results appears in Appendix E.

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

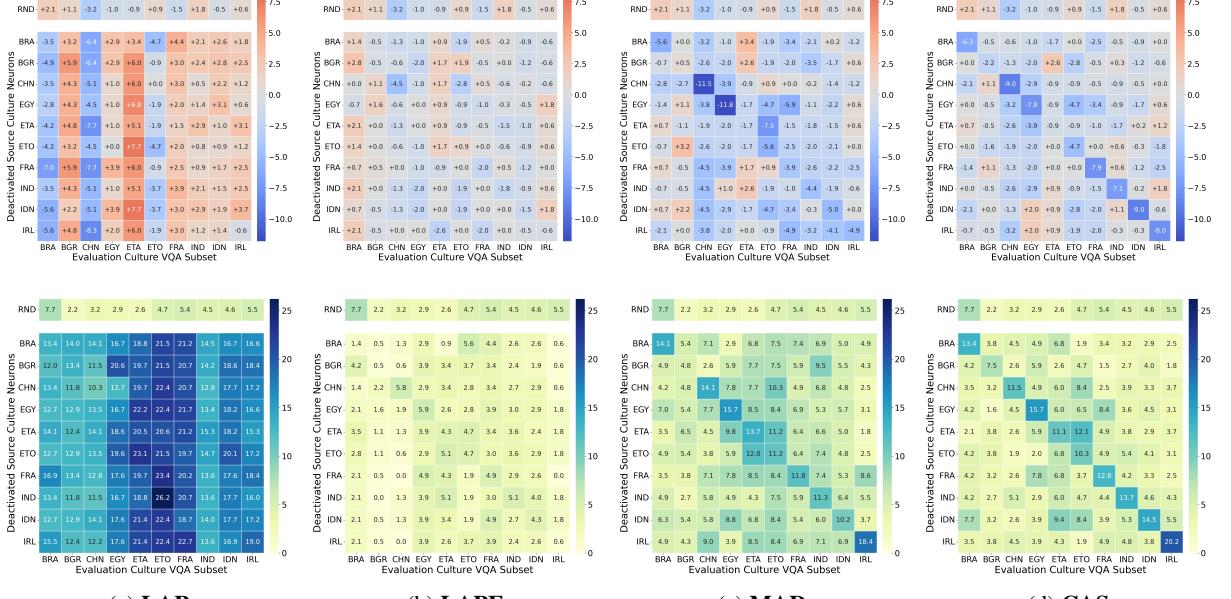


Figure 4: **Accuracy change** Δ (top) and **flip rate** heatmaps (bottom) on CVQA for different identification methods (Qwen2.5-VL-7B; showing first ten cultures). On the y-axis we have the source culture for which neurons are identified and ablated and on the x-axis the culture used for evaluation. We report percentage changes relative to the unablated full model. Diagonal cells show self-deactivation results. Results for all culture pairs are in Appendix D.3.

5 Results

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 (≈ 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 presents the main results. We report accuracy change (Δ) 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).

Qwen2.5-VL-7B and Pangea-7B. CAS yields the largest self-deactivation drops in accuracy

(Qwen: -5.52% ; Pangea: -4.33%) paired with small cross-deactivation changes ($< 1\%$), showing 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, while this seems counterintuitive, this is a known phenomenon that can be explained as pruning overly dominant or noisy activations (Ali et al., 2025).

LLaVA-v1.6-Mistral-7B. For LLaVA, LAP induces the strongest self-deactivation drop (-2.5%) but also larger cross-cultural spillover. CAS and MAD are more isolated yet smaller in magnitude, suggesting that cultural information is more diffusely distributed in this model.

5.2.1 Culture-Specific Patterns

We now delve into a more fine-grained per-culture analysis. Figure 4 visualizes accuracy changes (Δ) and flip-rates for Qwen2.5-VL-7B. For readability, we visualize the first ten cultures; see Appendix D for full results. We find that CAS produces sharp

diagonal degradations (e.g., CHN, FRA) with limited but non-negligible off-diagonal changes, evidencing strong mapping between masked neuron sets and cultures. Yet, LAP shows broad column-shaped reductions (large off-diagonals), pointing to less specific features. Interestingly, we find that for many cultures deactivating neurons associated with any culture positively impacts performance (e.g. BGR, ETA), suggesting that those neurons were negatively intervening. Specifically, we observe clear distinctions between cultures for which ablation always has a negative (e.g. BRA, CHN) or positive effect. Furthermore, LAPE reveals fairly little selectivity, while MAD sits between LAPE and CAS. The flip-rate matrices mirror these distinctions: CAS achieves the cleanest separation between self and cross conditions.

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 early-mid layers (6–8), with relatively sparse presence in deeper blocks. Interestingly, MAD tends to bypass the central layers (15–18), whereas CAS identifies neurons more evenly across mid-to-late layers. CAS 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 confirm that *instruction-following remains intact*: even after ablation, the model respects the instruction format constraint (i.e. generating one of the listed options), suggesting that decoder-level mask-

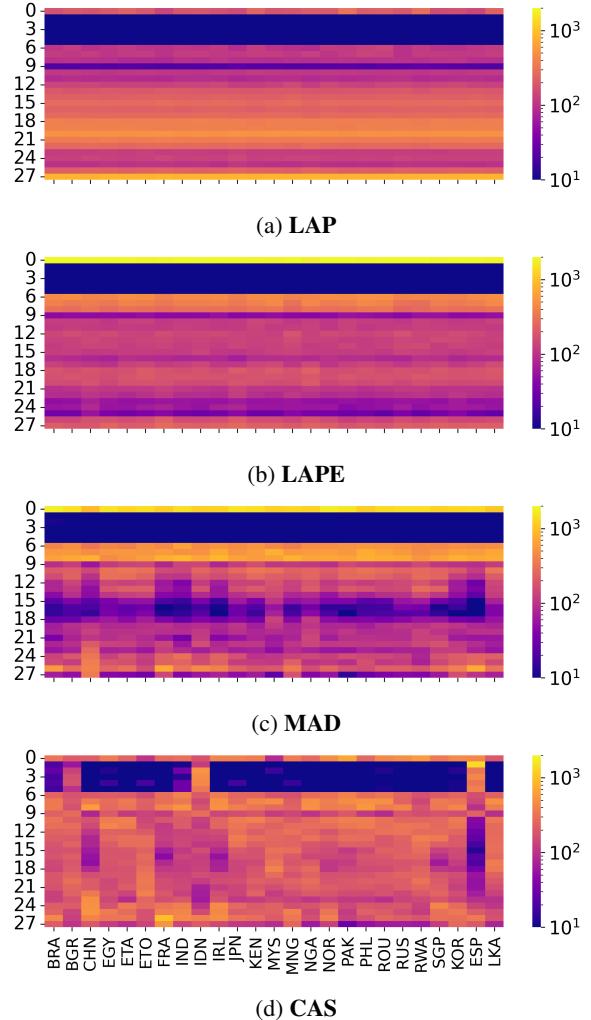


Figure 5: **Layer-wise counts of identified neurons** by different methods (Qwen2.5-VL-7B; log-scaled color).

ing does not broadly damage generation or task framing. 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 CAS 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 neu-

rons are not uniformly distributed but cluster in the foremost and early mid decoder layers. (4) Among all identification methods, CAS 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 VLM’s culture-specific performance. We introduce a margin-based selector (CAS) that allows for more precise identification of culture-sensitive neurons. Our method detects neurons whose ablation yields the largest self-deactivation drops with minimal cross-deactivation spillover on Qwen2.5-VL-7B and Pangea-7B. Layer-wise analyses reveal consistent concentration of culture-sensitive neurons across cultures. These results highlight the potential for small, targeted, activation interventions to mitigate cultural biases and improve cultural alignment without retraining the full model. Future work should extend the search beyond decoder MLPs and pair identification with activation steering.

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.

Limitations

Defining ‘culture’. We use CVQA’s country–language taxonomy and, for fairness to multilingual models, solely the English-translated prompts to 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 publicly available datasets (primarily 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.

References

- Ameen Ali, Shahar Katz, Lior Wolf, and Ivan Titov. 2025. Detecting and pruning prominent but detrimental neurons in large language models. *Preprint*, arXiv:2507.09185.
- Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. See it from my perspective: How language affects cultural bias in image understanding. *Preprint*, arXiv:2406.11665.
- Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulic. 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, Sibo 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. 2018. [Identifying and controlling important neurons in neural machine translation](#). *Preprint*, arXiv:1811.01157.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. [Network dissection: Quantifying interpretability of deep visual representations](#). *Preprint*, arXiv:1704.05796.
- 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.
- 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.
- Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. [Grounding multilingual multimodal llms with cultural knowledge](#). *Preprint*, arXiv:2508.07414.
- Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. [Towards neuron attributions in multi-modal large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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](#). *Preprint*, arXiv:2401.12181.
- 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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.
- 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](#). *Preprint*, arXiv:2301.12597.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Avinash Madasu, Vasudev Lal, and Phillip Howard. 2025. [Cultural awareness in vision-language models: A cross-country exploration](#). *Preprint*, arXiv:2505.20326.
- Danial Namazifard and Lukas Galke. 2025. Isolating culture neurons in multilingual large language models. *arXiv preprint arXiv:2508.02241*.
- 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](#). *Preprint*, arXiv:2410.07149.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghafari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.
- 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, Jiaxi 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, Ulisso 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, Xuejin Chen, and Yan Lu. 2025. [Deciphering functions of neurons in vision-language models](#). *Preprint*, arXiv:2502.18485.
- Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. [Beyond words: Exploring cultural value sensitivity in multimodal models](#). *Preprint*, arXiv:2502.14906.
- 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 lilm for 39 languages](#). *Preprint*, arXiv:2410.16153.
- 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. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

A Reproducibility

A.1 Models and Sources

Models	Sources
LLaVA-v1.6-Mistral-7B	https://huggingface.co/llava-hf/LLaVA-v1.6-Mistral-7B-hf
Pangea-7B	https://huggingface.co/neulab/Pangea-7B
Qwen2.5-VL-7B	https://huggingface.co/Qwen/qwen2.5-vl-7b-Instruct

Table 3: Specification and 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.

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. The complete procedure is summarized in Algorithm 1.

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

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.

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., 2023; 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 no choice can be confidently identified using this rule, the system falls back to a less robust substring search, checking whether the ground-truth option appears anywhere in the prediction text.

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.

Algorithm 1: Answer Normalization and Extraction

```

Input: Prediction string  $P$ , list of choices
 $C$ , ground truth  $G$ 
Output: Boolean indicating whether
prediction is correct
normalize  $P$ : lowercase, collapse
whitespace, trim edges;
Initialize  $last\_choice \leftarrow None$ ,
 $last\_pos \leftarrow -1$ ;
foreach choice  $c \in C$  do
    normalize  $c$ ;
    Find all word-boundary matches of  $c$  in
 $P$ ;
    foreach occurrence at position  $pos$  do
        if  $pos > last\_pos$  then
             $last\_pos \leftarrow pos$ ;
             $last\_choice \leftarrow c$ ;
if  $last\_choice \neq None$  then
    return ( $last\_choice = G$ );
else
    return (whether  $G$  appears in  $P$  as
substring);

```

B Activation Recording and Identification Details

B.1 Activation Extraction

We record neuron activations by attaching forward hooks to the nonlinearity branch of each decoder MLP. Specifically, in common transformer implementations, this branch is named `act_fn` in the SwiGLU block:

$$u = h_{l-1}W_u + b_u, \quad v = h_{l-1}W_v + b_v, \\ g = \text{SiLU}(u), \quad z = (g \odot v)W_o + b_o. \quad (1)$$

For each neuron n in layer l and token position t , we log the scalar activation $a_{l,n,t}$ from g . Because $\sigma(\cdot) > 0$, $\text{SiLU}(x)$ shares the sign of its pre-activation x , so $\mathbb{I}(a_{l,n,t} > 0)$ is equivalent to $\mathbb{I}(u_{l,n,t} > 0)$, making sign-based counts inexpensive.

We also ensure that the valid-token mask m_t respects each model’s internal attention mask, thereby excluding padding, image delimiters, and other special tokens.

B.2 Baseline Identification Method Implementation

B.2.1 Random Selection (RND)

We use a *global* random baseline that samples a fixed total number of neurons (i.e. $r\%$) as the targeted mask but draws them uniformly from *all* layers, without enforcing any layer-wise quota. This choice avoids additional bookkeeping and is substantially more compute-efficient than a layer-matched variant that mirrors a method’s per-layer counts. To verify that conclusions are not an artifact of the layer distribution, we probed a *layer-matched* random baseline (matching CAS’s per-layer histogram). Accuracy changes and flip rates were comparable to the global RND within sampling variance, so we report the global RND in the main text for simplicity and efficiency.

B.2.2 Activation Probability (LAP)

LAP (Gurnee et al., 2024; Voita et al., 2024) selects neurons that often fire for a given aspect (‘language’ in the original works; here adapted to culture). Using the activation probability from Eq. (2),

$$P_{l,n}^{(c)} = \frac{1}{T_c} \sum_{t=1}^{T_c} \mathbb{I}(a_{l,n,t}^{(c)} > 0) \quad (2)$$

we define the LAP score for culture c as

$$s_{l,n}^{\text{LAP}}(c) = P_{l,n}^{(c)}. \quad (3)$$

We apply two simple filters before selection:

- Activity filter.** Compute a global activity threshold p_{th} as the α -percentile over all values in $\{P_{l,n}^{(c)}\}$ (we use $\alpha=95$). Discard triples (l, n, c) with $P_{l,n}^{(c)} < p_{\text{th}}$ to remove rarely firing neurons.
- Competition tie-break.** When needed, prefer neurons with a larger probability margin $\delta_{l,n}^{\text{LAP}}(c) = P_{l,n}^{(c)} - \max_{c' \neq c} P_{l,n}^{(c')}$, encouraging comparative culture selectivity without computing entropy.

For each culture c , we rank the remaining neuron indices (l, n) by $s_{l,n}^{\text{LAP}}(c)$ (with $\delta_{l,n}^{\text{LAP}}(c)$ as a tie-breaker) and then select the top $r\%$ across all decoder MLP neurons (we use $r=1$).

B.2.3 Activation Probability Entropy (LAPE)

LAPE (Tang et al., 2024; Namazifard and Galke, 2025) prefers neurons whose activations across cultures are *concentrated*. Using the same activation probabilities $P_{l,n}^{(c)}$, define the normalized distribution

$$\tilde{p}_{l,n}^{(c)} = \frac{P_{l,n}^{(c)}}{\sum_{c'} P_{l,n}^{(c')}},$$

and compute the Shannon entropy

$$H_{l,n} = - \sum_{c \in \mathcal{C}} \tilde{p}_{l,n}^{(c)} \log \tilde{p}_{l,n}^{(c)}. \quad (4)$$

Lower $H_{l,n}$ indicates stronger specialization.

- Activity filter.** Drop (l, n) with $\max_c P_{l,n}^{(c)} < p_{\text{th}}$, using the same global activity threshold p_{th} (the α -percentile over all $\{P_{l,n}^{(c)}\}$; we use $\alpha=95$).
- Low-entropy pool.** Among the survivors, keep a low-entropy pool by taking the lowest $\rho\%$ according to $H_{l,n}$ (chosen only to form a sufficiently selective candidate set).
- Culture assignment.** Assign each kept neuron (l, n) to its top culture $c^* = \arg \max_c P_{l,n}^{(c)}$, provided $P_{l,n}^{(c^*)} \geq p_{\text{bar}}$, where p_{bar} is the β -percentile of all $\{P_{l,n}^{(c)}\}$ (we use $\beta=90$).

For each culture c , rank the neurons assigned to c by decreasing $P_{l,n}^{(c)}$ (breaking ties by lower $H_{l,n}$, then by larger margin $P_{l,n}^{(c)} - \max_{c' \neq c} P_{l,n}^{(c')}$). Finally, select the top $r\%$ across all decoder MLP neurons for that culture.

B.2.4 Mean Activation Difference (MAD)

MAD scores neurons by how much more strongly they respond to culture c than to other cultures, using magnitude-aware signals (Bau et al., 2018; Dalvi et al., 2019). Using the mean positive activation $M_{l,n}^{(c)}$ from Eq. (1), define

$$\bar{M}_{l,n}^{(-c)} = \frac{1}{|\mathcal{C}| - 1} \sum_{c' \in \mathcal{C} \setminus \{c\}} M_{l,n}^{(c')}, \quad (5)$$

$$s_{l,n}^{\text{MAD}}(c) = M_{l,n}^{(c)} - \bar{M}_{l,n}^{(-c)}. \quad (6)$$

To avoid selecting neurons that spike rarely but strongly, we apply the same activity gate as in LAP/LAPE:

$$\text{keep } (l, n, c) \text{ only if } P_{l,n}^{(c)} \geq p_{\text{th}}, \quad (7)$$

where p_{th} is the α -percentile over all $\{P_{l,n}^{(c)}\}$ (we use $\alpha=95$). For scale stability across layers, one may optionally apply a layer-wise z -normalization to M before computing (6); our main results use the unnormalized form with the gate in (7).

For each culture c , rank all decoder-MLP neurons (l, n) by decreasing $s_{l,n}^{\text{MAD}}(c)$. Break ties by larger $P_{l,n}^{(c)}$, then by larger margin $M_{l,n}^{(c)} - \max_{c' \neq c} M_{l,n}^{(c')}$. Finally, select the top $r\%$ across all decoder MLP neurons for that culture.

B.3 Metrics

Let N be the number of items in a culture subset. Let \hat{a}_i^{mask} and \hat{a}_i^{full} denote the model’s predicted answers with and without ablation masking (after normalization; §A.4) for item i before and after masking, respectively. Let $y_i^{\text{full}}, y_i^{\text{mask}} \in \{0, 1\}$ indicate correctness under the two runs.

$$\text{Acc}_{\text{full}} = \frac{1}{N} \sum_{i=1}^N y_i^{\text{full}}, \quad (8)$$

$$\text{Acc}_{\text{mask}} = \frac{1}{N} \sum_{i=1}^N y_i^{\text{mask}}, \quad (9)$$

$$\Delta \text{Acc} = \text{Acc}_{\text{mask}} - \text{Acc}_{\text{full}}. \quad (10)$$

Flip rate. We measure the proportion of items whose predicted answers change:

$$\text{FlipRate}_{\text{any}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{a}_i^{\text{orig}} \neq \hat{a}_i^{\text{mask}}). \quad (11)$$

C Layer-wise Neuron Distribution

We illustrate the number of identified neurons per layer across all selected cultures of LLaVA-v1.6-Mistral-7B in Figure 6 and Pangea-7B in Figure 7. We use a logarithmic scale to compress the color range for the very high values in dominant layers and expand the color range for the lower values in the other layers.

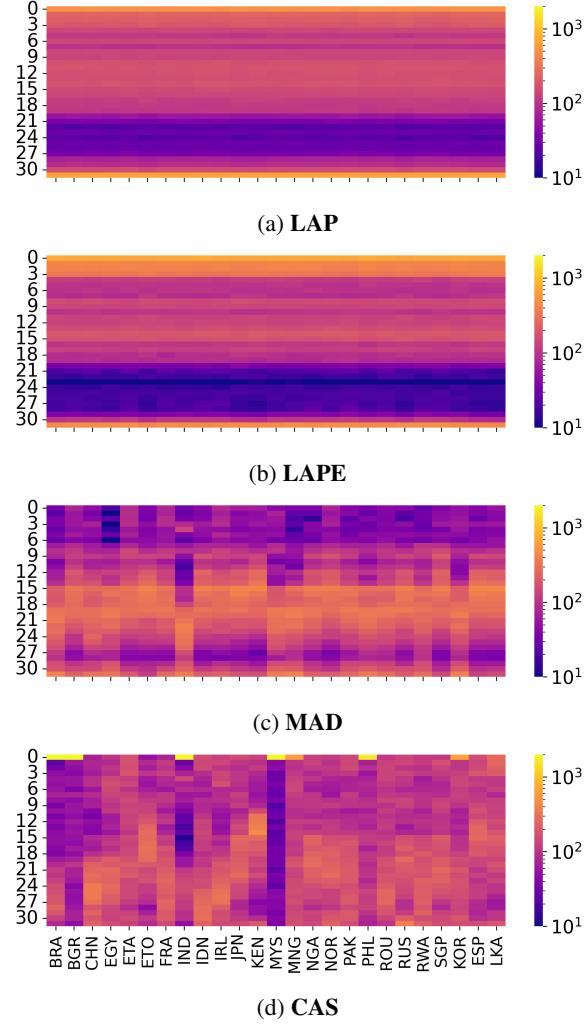


Figure 6: Layer-wise counts of identified neurons (LLaVA-v1.6-Mistral-7B).

LLaVA-v1.6-Mistral-7B Culture-sensitive neuron distributions vary more across methods than Qwen2.5-VL-7B and Pangea-7B. LAP and LAPE emphasize first and last layers. MAD shows concentrations in mid-to-late layers (15–24). CAS is sparser than MAD and locates more neurons in early layers (0–6).

Across early-to-mid layers (1–20), both methods also have coverage, whereas the clear purple band between layers (21–27) indicates neglect of those layers. MAD shows concentrations in mid-to-late layers (15–24). CAS is sparser than MAD and locates more neurons in early layers (0–6).

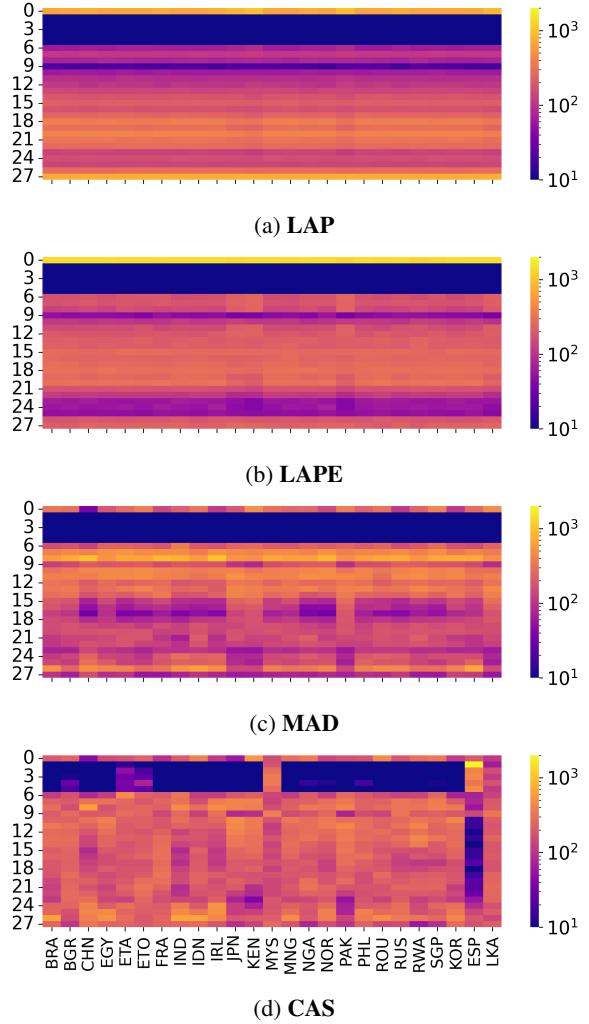


Figure 7: Layer-wise counts of identified neurons (Pangea-7B).

Pangea-7B Activation density distribution across layers for Pangea-7B is very similar to the one for Qwen2.5-VL-7B, as shown in Figure 7. This is largely expected, as the language component of Pangea-7B uses a Qwen2.7B-Instruct backbone (Yang et al., 2024), which is the direct predecessor of Qwen2.5-7B-Instruct (Qwen et al., 2025), the language component of Qwen2.5-VL-7B. They both share a 28-layer architecture. Similar with Qwen2.5-VL-7B, the aggregated culture ESP exhibits distinctive patterns in comparison with other cultures, with neuron concentration on early layers (0–5).

D Full Ablation Matrices

D.1 Pangea-7B Ablation Results

Accuracy change. In Figure 8, off-diagonal patterns reveal key differences in cultural specificity between methods. LAP produces widespread cross-

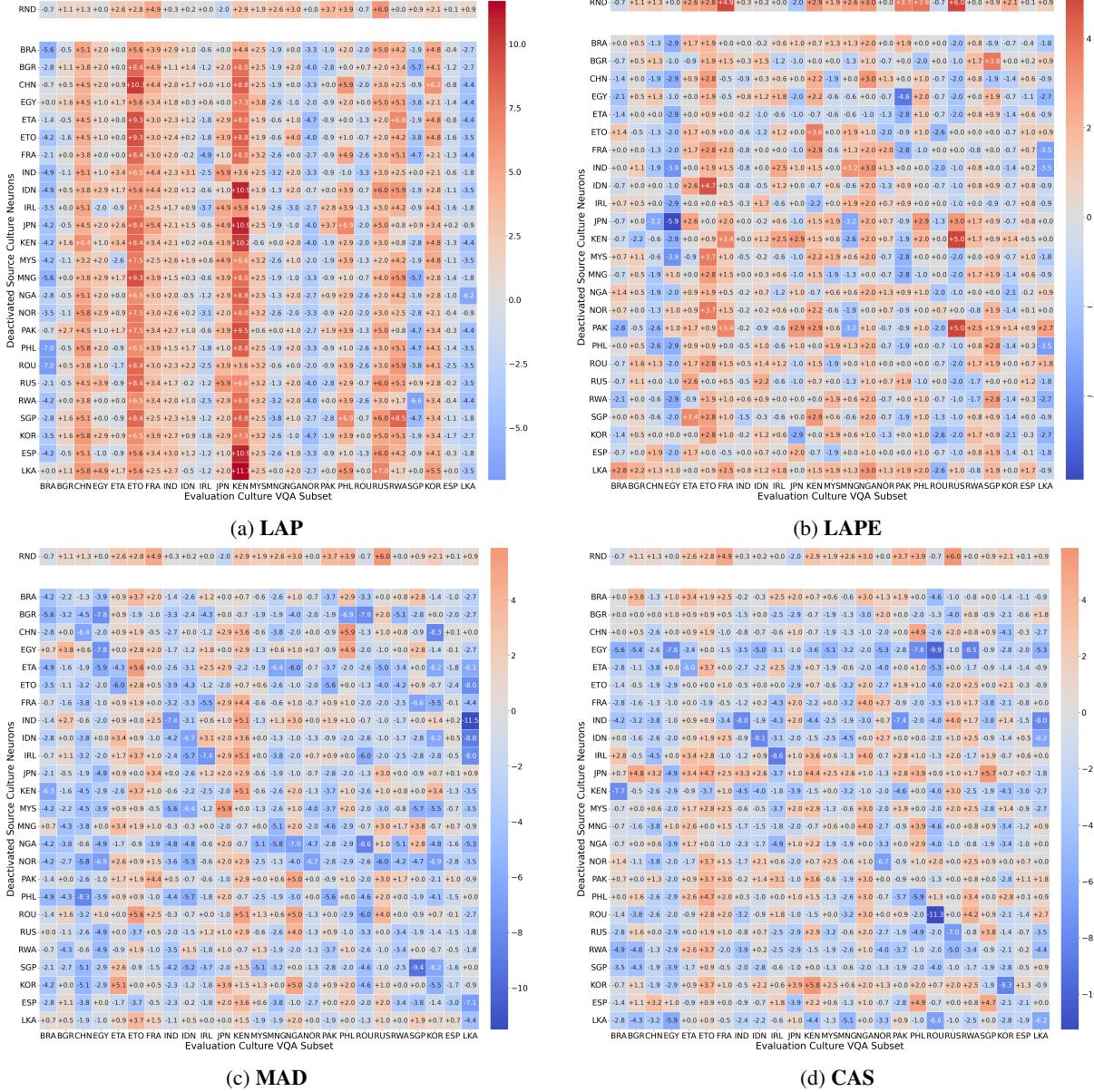


Figure 8: Accuracy change Δ heatmap for Pangea-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.

cultural interference, with off-diagonal values vary dramatically (exceeding -5% and $+10\%$). This suggests that LAP often selects neurons encoding features shared across multiple cultures, potentially diluting cultural isolation. LAPE reduces this interference, with most off-diagonal values between -2% and $+2\%$, but its self-deactivation magnitudes are generally smaller, implying weaker capture of high-impact culture-sensitive neurons.

MAD and CAS produce the largest and most consistent negative accuracy changes in the self-deactivation diagonal. MAD exhibits particularly strong effects for SGP (-9.4%) and EGY (-7.8%), indicating that its selected neurons are highly influ-

ential for these cultural subsets. CAS yields large drops for ROU (-11.3%) and IRL (-8.6%).

MAD, while producing strong self-deactivation drops, also shows notable cross-effects for related cultural groups, especially within African and Asian subsets, indicating partial overlap in feature encoding. In contrast, CAS achieves a favorable balance: its self-deactivation effects are large, yet most cross-deactivation changes remain within $\pm 1.5\%$, demonstrating targeted neuron removal that minimally impacts unrelated cultural subsets.

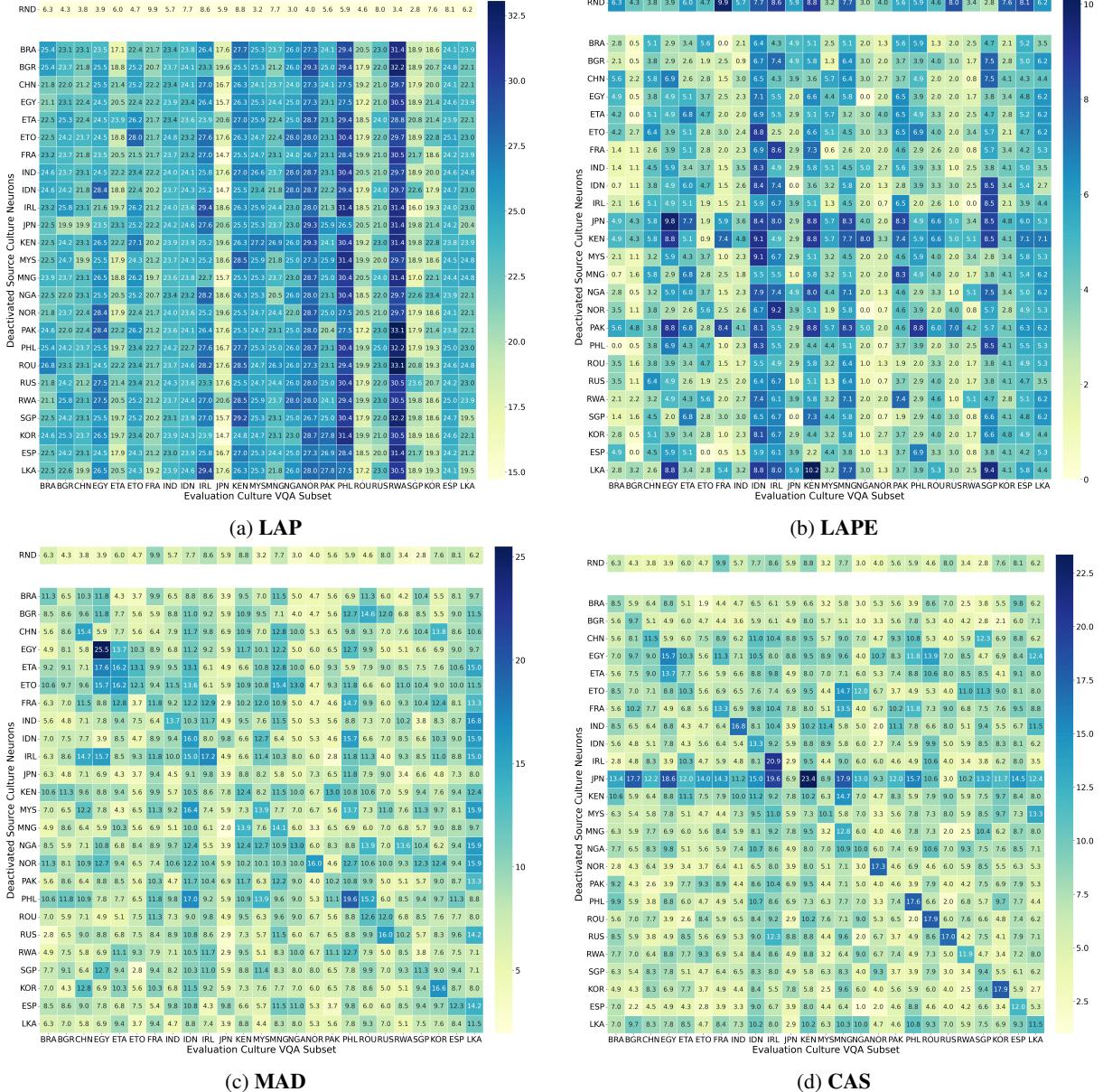


Figure 9: **Flip Rate heatmap for Pangea-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.**

Flip rate. In Figure 9, LAP produces the highest and most widespread flip rates, with substantial off-diagonal deactivation values exceeding 25% (e.g., EGY, RWA, PHL, etc.) LAPE markedly reduces off-diagonal rates (generally below 7%) but also yields lower self-deactivation magnitudes.

MAD achieves large self-deactivation flip rates (often 10 – 20%) but also moderate cross-effects, especially among related cultural groups. CAS offers a strong balance: self-deactivation rates remain high for several cultures (e.g., 20.9% for IRL, 17.9% for ROU and KOR), while off-diagonal effects are generally modest, indicating effective cultural specificity. We notice the unique pattern

shown for deactivating JPN neurons. After dissecting the results, we found that this was likely due to the model’s initial bias (extreme low performance on JPN-specific training subset), causing the models fail to accumulate culture-sensitive activation signals through sufficient successful VQA cases.

Overall, these patterns align with the accuracy Δ results. LAP fail to target culture-sensitive neurons in general; LAPE improves the isolation of such neurons but the results remain unsatisfactory; MAD captures powerful but partially shared representations, and CAS combines substantial self-impact with minimal cross-cultural interference.

Taken together, for Pangea-7B, the heatmaps re-

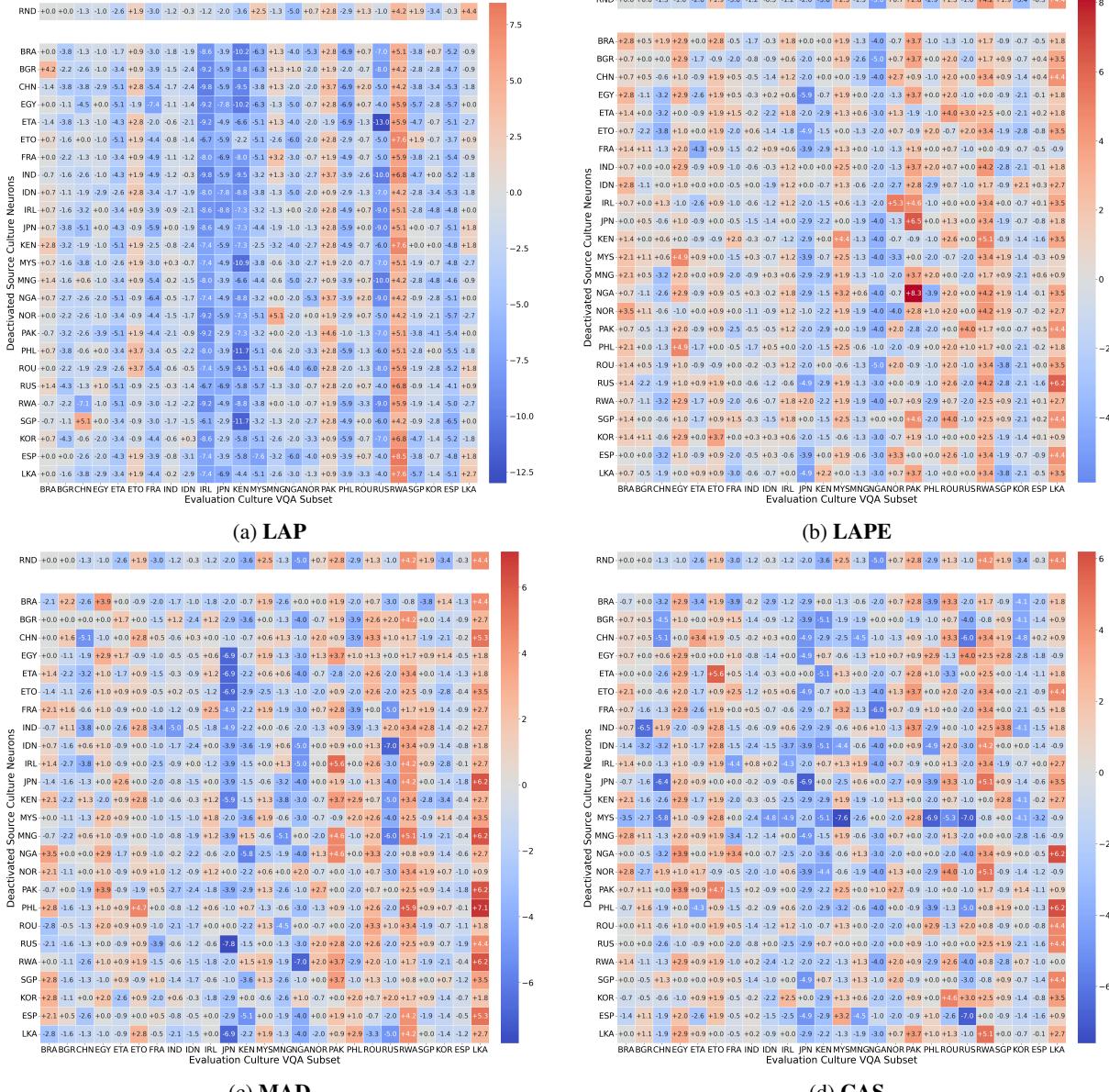
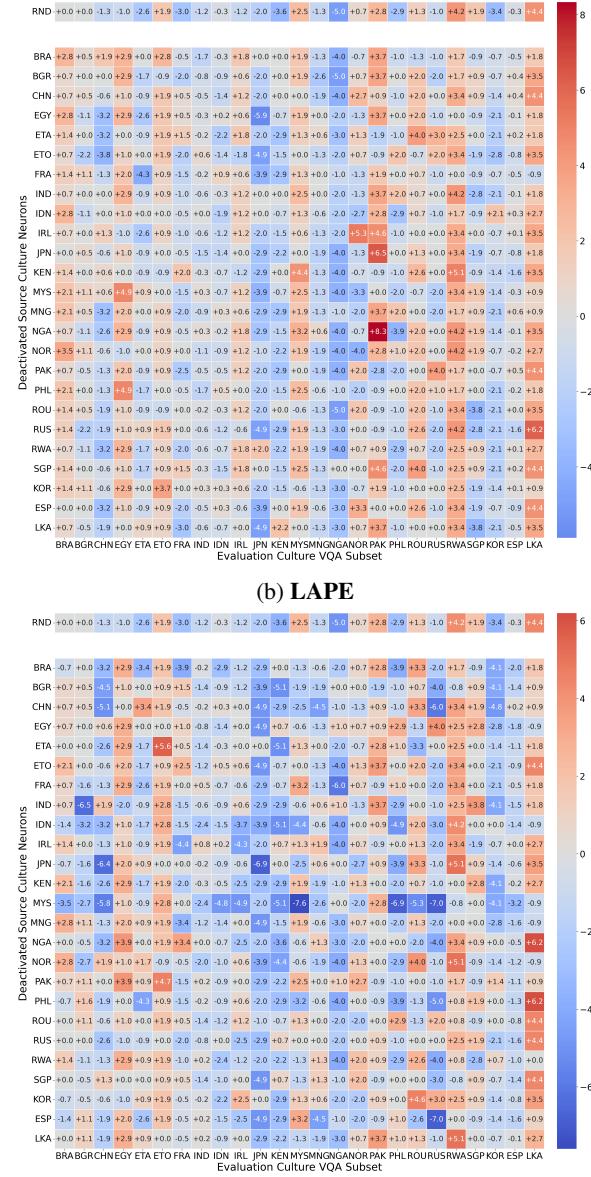


Figure 10: Accuracy change Δ heatmap for LLaVA-v1.6-Mistral-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.

inforce CAS’s strength in isolating neurons with concentrated cultural influence while maintaining low collateral impact. LAP shows broader but less specific disruption, LAPE favors specificity at the expense of impact, and MAD captures strong but partially entangled cultural representations. These results are consistent with the patterns observed for Qwen2.5-VL-7B, suggesting that CAS’s advantages generalize across models with different training data and architecture choices.

D.2 LLaVA-v1.6-Mistral-7B Ablation Results

Accuracy change. As demonstrated in Figure 10, LAP induces the largest and most consistent ac-



accuracy change variance drops (between -10% and $+5\%$), but show no discrimination on self-deactivation effects and cross-deactivation interferences. This aligns with the aggregate results in Table 2, where LAP yields the strongest self-deactivation impact for this model. LAPE shows similar incapability of distinguishing culture-sensitive neurons but with smaller impact magnitude. In contrast, CAS and MAD produce stronger diagonal drops than cross-diagonal ones on average, suggesting that the neurons identified by these methods are more likely to contribute to culture-sensitive information processing.

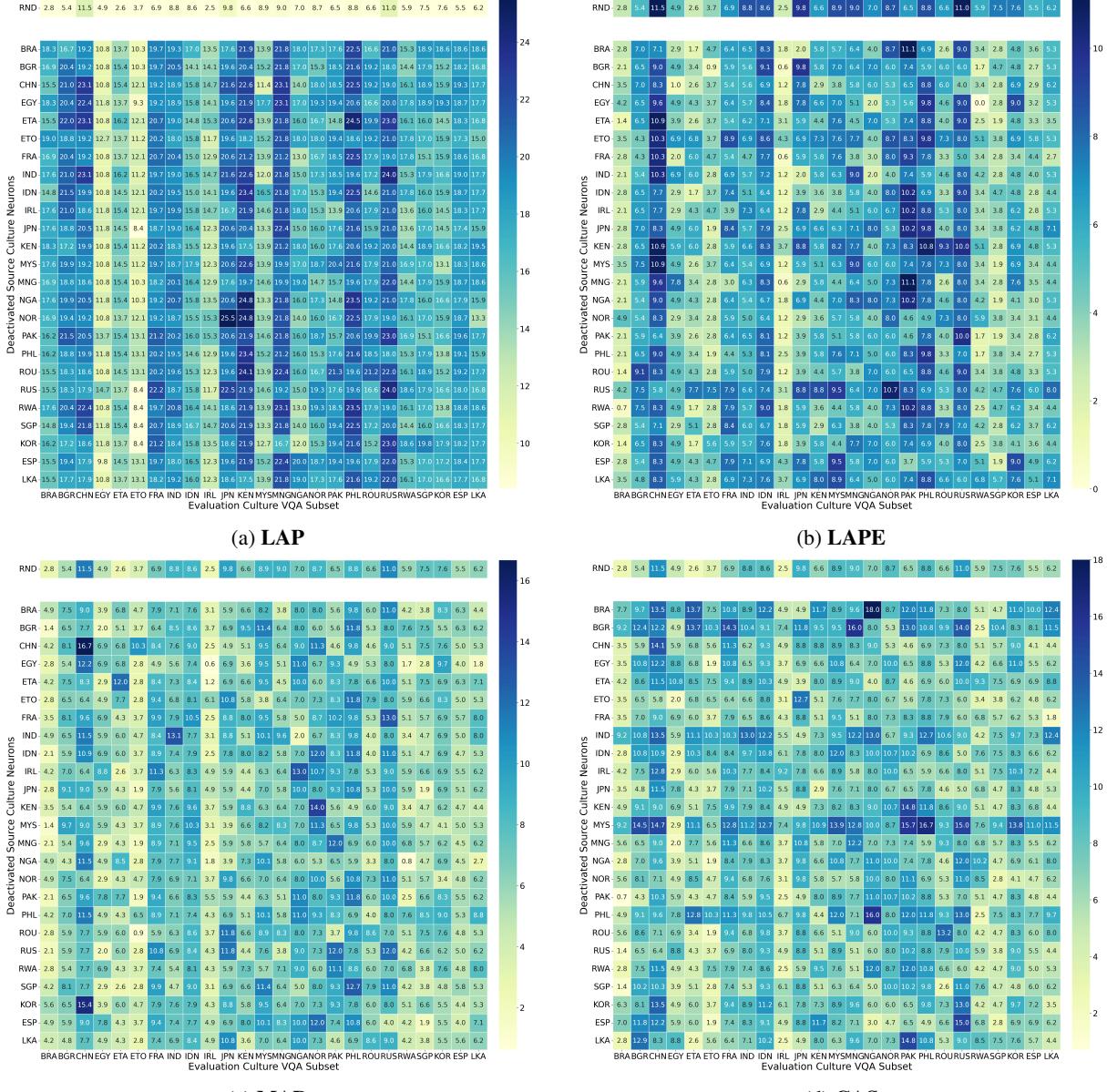


Figure 11: **Flip Rate** heatmap for LLaVA-v1.6-Mistral-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.

Flip rate. Figure 11 shows the flip rates for LLaVA-v1.6-Mistral-7B. LAP yields the highest self-deactivation flip rates, with multiple diagonal values exceeding 30% and substantial off-diagonal effects across unrelated cultures. LAPE reduces off-diagonal interference, generally keeping cross-cultural flip rates below 10%, but also produces smaller self-deactivation rates (also often under 10%).

MAD results in moderate self-deactivation flip rates (7-13%) and low-to-moderate cross-effects, pointing to a balance between specificity and impact, though less pronounced than in Qwen2.5-VL-7B or Pangea-7B. CAS produces mixed results for

this model. Self-deactivation flip rates are moderate (10-20% for the most affected cultures), and off-diagonal values remain relatively contained. While CAS maintains reasonable specificity, its self-impact here is smaller than for the other two models.

Overall, for LLaVA-v1.6-Mistral-7B, LAP emerges as the most disruptive method in terms of raw self-deactivation impact, albeit with higher cross-cultural spillover, while LAPE produces a mixed pattern of moderate self-impact and dispersed cross-effects. CAS and MAD maintain higher specificity with smaller cross-deactivation effects. These results could suggest that cultural

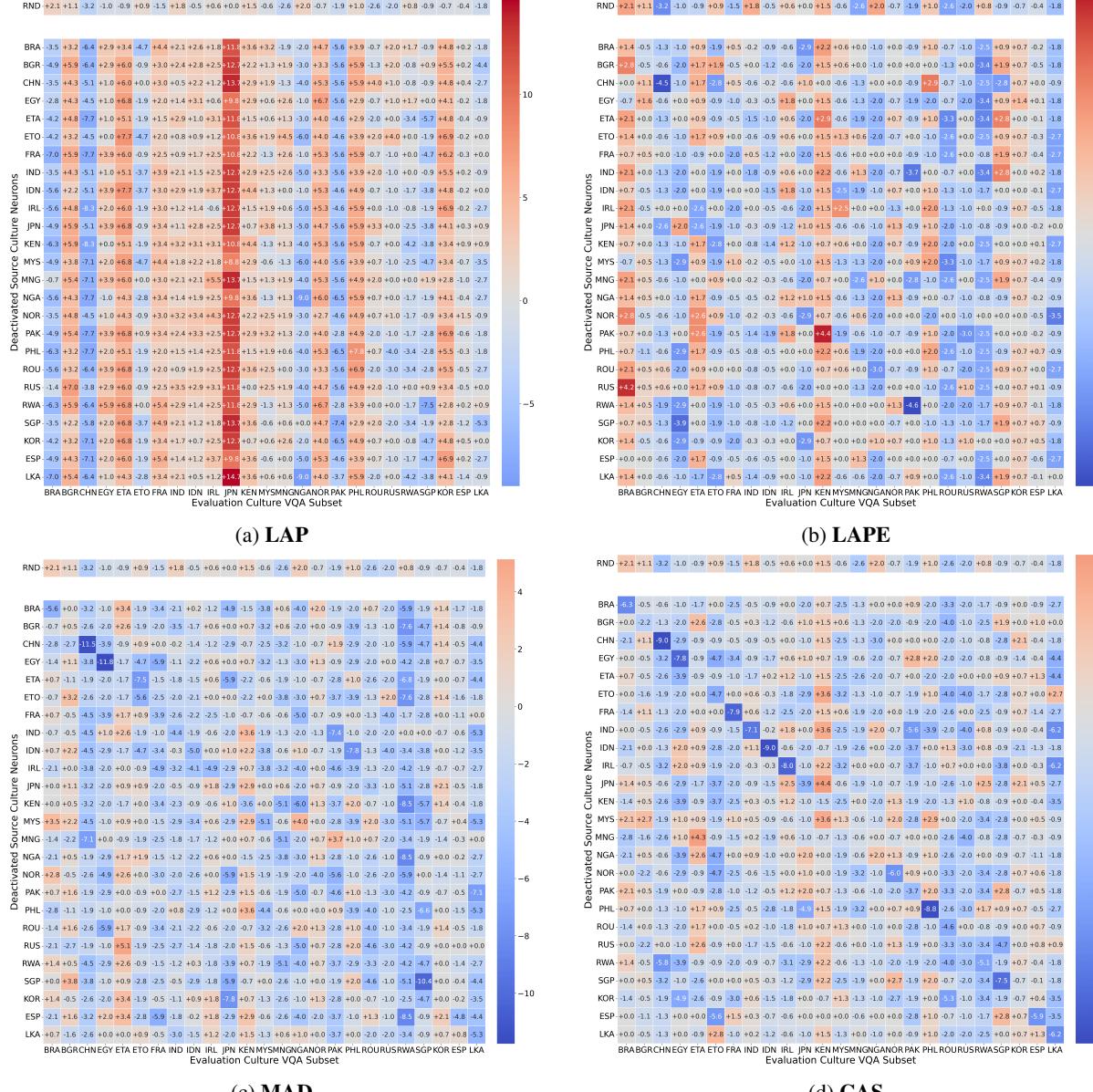


Figure 12: Accuracy change Δ heatmap for Qwen2.5-VL-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.

knowledge in this model is more diffusely encoded, making high-specificity neuron targeting less impactful than in Qwen2.5-VL-7B or Pangea-7B.

D.3 Qwen2.5-VL-7B Ablation Results

Full results are presented in Figures 12 and 13.

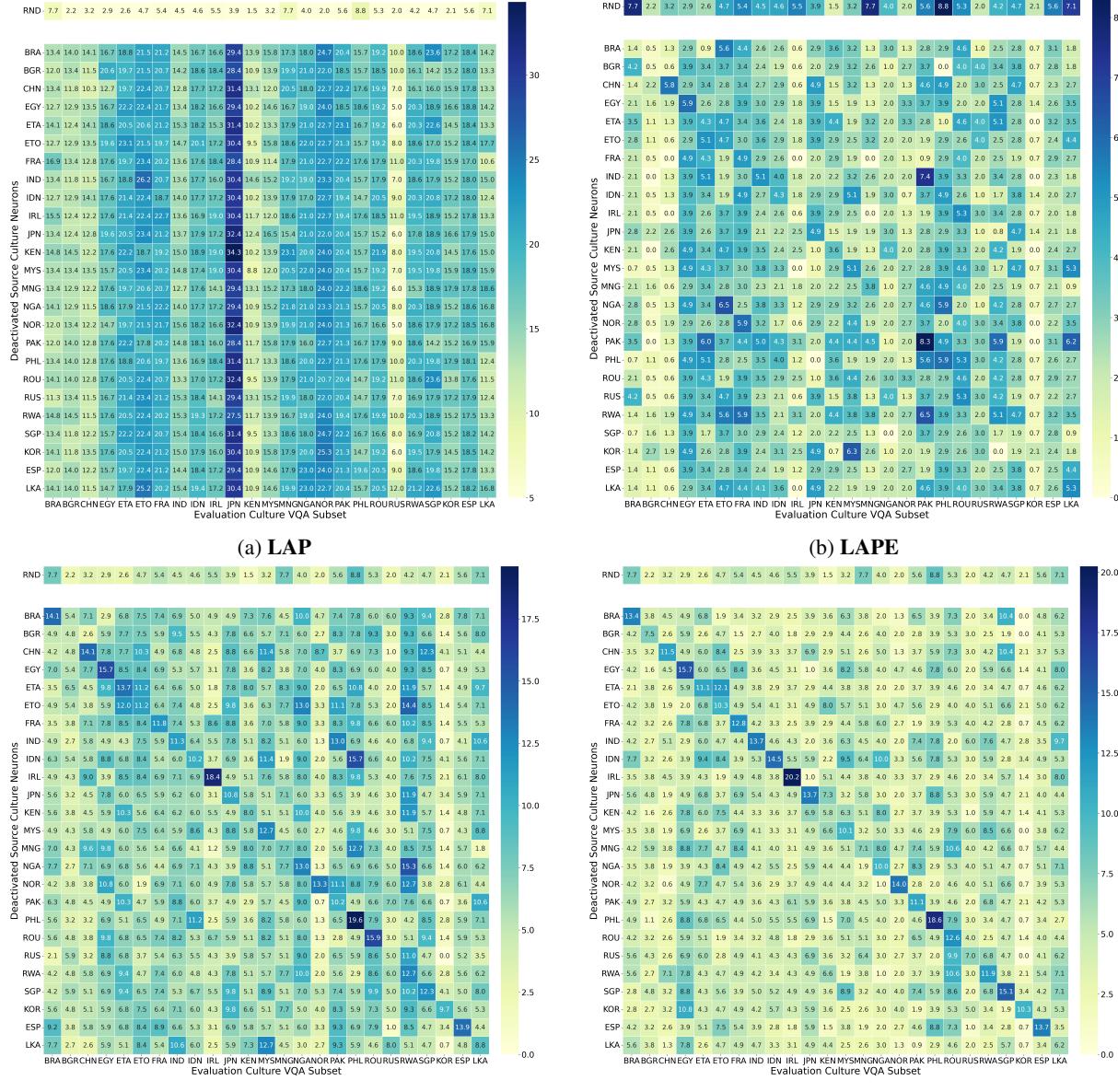


Figure 13: **Flip Rate** heatmap for Qwen2.5-VL-7B after deactivating neurons identified by different methods, shown for all evaluated cultures.

E 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 model 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	65101 (12.27%)	50772 (9.57%)	148 (0.03%)

Table 6: **Neuron-wise standard deviations across cultures** for the three evaluated models. The preliminarily experiment on Qwen2.5-VL-7B and Pangea-7B yield high variances across cultures, which motivates us to develop CAS.