Vision Language Models are Confused Tourists

Patrick Amadeus Irawan[1*], Ikhlasul Akmal Hanif[1*], Muhammad Dehan Al Kautsar[1], Genta Indra Winata[2†], Fajri Koto[1†], Alham Fikri Aji[1†]
[1]MBZUAI [2]Capital One
[*]Main Authors, [†]Senior Authors
(patrick.irawan, ikhlasul.hanif)@mbzuai.ac.ae

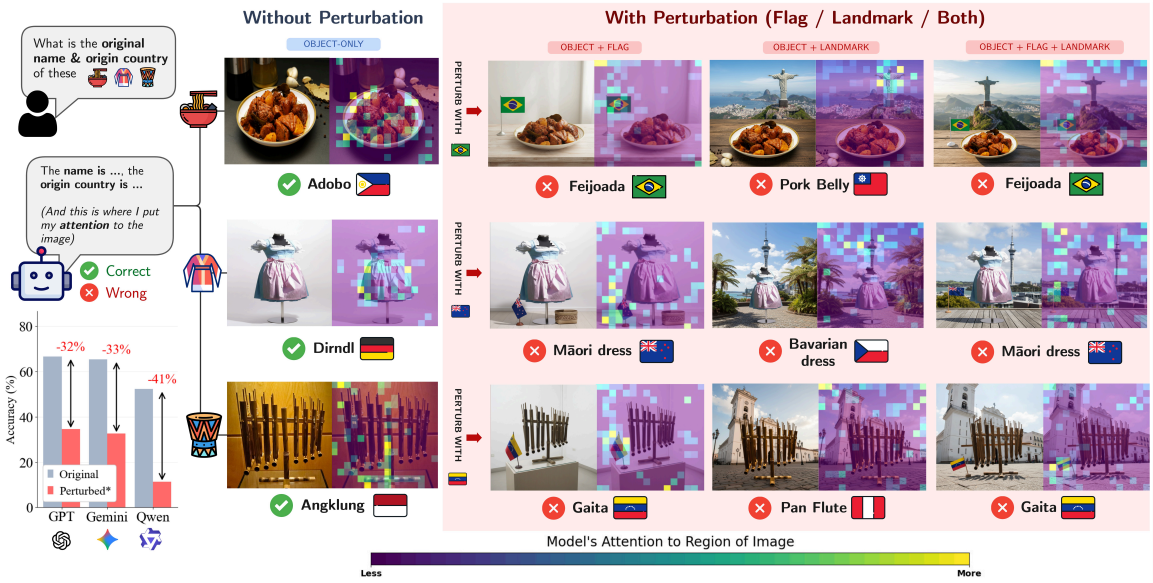


Figure 1. CONFUSEDTOURIST evaluates the robustness of current state-of-the-art VLMs on grounding single cultural concept within images that contain distracting cue(s). We demonstrate how geographical-induced perturbation causes massive accuracy drop (flag-perturbation*) consistently across all cases. Through further interpretability analysis, we also discover that the model's distractive attention shift toward the adversarial cue(s) (e.g., Row 1 example where the Brazil flag was attended significantly more than the Adobo cuisine) can directly explain the decline.

Abstract

Although the cultural dimension has been one of the key aspects in evaluating Vision-Language Models (VLMs), their ability to remain stable across diverse cultural inputs remains largely untested, despite being crucial to support diversity and multicultural societies. Existing evaluations often rely on benchmarks featuring only a singular cultural concept per image, overlooking scenarios where multiple, potentially unrelated cultural cues coexist. To address this gap, we introduce CONFUSEDTOURIST *, a novel cultural adversarial robustness suite designed to assess VLMs' stability against perturbed geographical cues. Our experiments reveal a critical vulnerability, where accuracy drops heavily under simple image-stacking perturbations and even worsens with its image-generation-based variant. Interpretability analyses further show that these failures stem from systematic attention shifts toward distracting cues, diverting the model from its intended focus. These findings highlight a critical challenge: visual cultural concept mixing can substantially impair even state-of-the-art VLMs, underscoring the urgent need for more culturally robust multimodal understanding.* [1][2]

[1]Code: https://github.com/patrickamadeus/vlms-are-confused-tourists
[2]Data: https://huggingface.co/datasets/patrickamadeus/vlms-are-confused-tourists

1. Introduction

Recent advances in multimodality have enabled Vision-Language Models (VLMs) to become more proficient across a range of tasks, including ones that require domain-specific knowledge. Multicultural domain fits naturally into such description, as it requires models to have a culturally specific understanding to be able to capture relevant insight from rich visual inputs. Prior benchmarks have set the groundwork for such evaluation, including in the general-purpose scope [14, 19–21] and in a more fine-grained category setting (e.g. cuisine-only [11, 26] or paintings-only [28]). These systematic large-scale benchmarks were built to assess whether VLMs can probe and reason about multicultural knowledge beyond basic visual perception.

Although recent numbers[3] indicate collective improvement of recent VLMs' multicultural comprehension ability, aforementioned benchmarks only assess this understanding using unambiguous cultural images. These images typically contain either only a single concept (e.g., a person wearing a Japanese kimono) or multiple but inherently related cues (e.g., Indonesian Gamelan musicians playing drums and gongs next to temple dancers), thereby allowing grounding inference to be assisted by the existence of related cues that may not refer to the intended concept. This limitation makes it difficult to disentangle whether the VLM is identifying an object based on its intrinsic visual features or if it is overly relying on other contextual cues. Hence, the ideal setting for stress-testing this robustness is by involving scenes with simultaneously contrasting cultural cues to challenge the model's ability to refer to relevant object.

Prior works have attempted similar perturbation ideas in the multicultural domain. For instance, Ye et al. [27] conducted a text-modality perturbation check on cross-modal attention divergence in a cross-lingual setting, indicating that image understanding levels highly differ between languages. On the other modality side, Kim et al. [10] attempted to perturb the image by altering the ethnicity of persons to point out potential bias of VLMs. However, these works suffer from one or more of the following limitations: **(1)** primarily involve adversity just in text, **(2)** involve concepts that are inherently subjective or culturally biased, making explicit and objective analysis difficult and potentially leading to uneven treatment of different concepts (for example: racial or cultural groups in Kim et al. [10]), and **(3)** did not include a more comprehensive analysis over the model's failure behavior.

In this work, we address these critical gaps by introducing a novel evaluation suite designed to probe whether state-of-the-art VLMs are able to identify a specific cultural item amidst the existence of other perturbing cues. To achieve this, we construct a suite of test images where a target cultural item is deliberately co-presented with a conflicting geographical symbol or object (such as a flag or landmark). We include rich multicultural nuances across various cultural domains, concepts, perturbation contexts, image creation methods, together with empirical and behavioral analyses of notable failure cases. Our contributions are summarized as follows:

1. We present CONFUSEDTOURIST , a VL robustness evaluation suite comprising of 5k+ geographical-cue perturbed images which includes 243 unique culture items from 57 countries. We curate the altered images by applying 3 perturbation settings (flag-only, landmark-only, or both) on multiple difficulties.
2. We benchmark 14 SOTA VLMs on our suite, where we observe a consistent and critical performance degradation in all cases, even showing failure against the simple perturbing method (3.3.1).
3. We provide further interpretability analyses of open-source VLMs, highlighting their attention-shift behavior, biases, and reasoning thought to explain the nature of this performance drop.

2. Related Work

Multicultural Understanding in Vision Language. Prior work has shown that VLMs exhibit cultural blind spots due to training data heavily favoring Western contexts. Recent multicultural VLMs benchmarks [12, 19, 22, 26] consistently demonstrate that VLMs struggle to accurately recognize artifacts, foods, and traditions from non-Western or underrepresented regions. Studies such as Ye et al. [27] examine how text-modality perturbations affect cross-modal attention in cross-lingual settings, while Kim et al. [10] reveal that VLM predictions can be improperly influenced by irrelevant cues, such as the perceived ethnicity of a person in the image. Other works further confirm these challenges in multicultural grounding [14, 20]. Our research extends this line of inquiry by examining how a VLM's cultural understanding behaves under conflicting visual cues, addressing a critical and previously unexplored aspect of multicultural robustness.

VLM Robustness in Concept-Mixed Setting. VLM robustness research typically addresses two failure categories: failures against simple noise and failures against semantic complexity. Traditional studies focus on out-of-distribution corruption, such as added noise or common visual distortions [7, 8], which test models' basic perceptual stability [2, 3, 18, 23]. However, VLMs suffer from distinct vulnerabilities related to higher-level meaning and concept mixing. Failures have been observed in geometric reasoning [17], compositional generalization [2, 9, 15], and bias driven by

[3] We evaluated recent frontier VLMs to verify this observation; detailed results are provided in Appendix B.

Category	Ori.	Desc.		Geo.		Total Pairs
		Easy	Hard	Easy	Hard	
Cuisine (181)	181	894	1038	912	912	3756
Attire (38)	38	228	228	216	216	888
Music (24)	24	144	144	138	138	564
Grand Total	243	1266	1410	1266	1266	5451

Table 1. Dataset statistics for CONFUSEDTOURIST. With description- or geographical-based pairs, each image is perturbed at 2 difficulty levels using image stacking or generative perturbation. A more detailed stat of the suite can be seen in the Appendix.

strong visual or linguistic priors [23]. Furthermore, models frequently misattribute facts and hallucinate plausible but incorrect information, demonstrating weak factual grounding [17]. While prior work exists in multicultural settings [10, 27], these efforts fail to convey a detailed robustness study, especially one involving contrasting and non-subjective cultural concepts. Our study bridges this gap by treating cross-cultural visual mixing as a form of semantically guided perturbation, analyzing how VLMs fail to maintain factual focus when multiple, conflicting cultural concepts coexist.

3. CONFUSEDTOURIST Evaluation Suite

Our suite is comprised of 5,451 unique images, featuring 243 unique cultural items sourced from 57 countries across 11 sub-regions. The suite cover 3 categories, including cuisine, traditional attire, and musical instruments. We curate this suite using a 3-staged pipeline involving context crawling, pair creation, and image generation, with details depicted in Figure 2. Table 1 presents the overall statistics of our data.

3.1. Context Crawling

Cultural Domain. We select cuisine, traditional attire, and musical instruments as our categories because they meet two criteria: they are well-constrained and their items are typically represented by a single, clear, object-based cue. This approach allows us to reduce the ambiguity found in multi-concept categories (like festivals and games) or overly broad categories where items take many forms (like artifacts or gifts).

Geographical Context. Our country selection covers 57 countries sampled from 11 sub-regions, with each sub-region containing a maximum of 7 countries. We design this sampling strategy to balance sub-regional diversity, rather than focusing on larger divisions like continents. For each country's visual perturbation image grounding resources, we obtained flag and landmark images from Wikimedia Commons with a clear license record, detailed in Appendix A.

Cultural Item Pool. We select up to 5 cultural items for each category in each country, totaling 243 unique items. We extract only licensed images and their summarized item descriptions from Wikipedia or Wikimedia Commons. We further ensure their validity through a two-step quality check: (1) We cross-check the collected data against prior benchmarks [19, 20, 26]; and (2) We employ internal quality control by conducting a blind, swapped peer review among all authors for every instance. Finally, each cultural item includes its name, country, description, and its corresponding image.

3.2. Adversarial Pairing.

We employ 2 pairing methods following cultural proxies introduced by Adilazuarda et al. [1]. First, description-based pairing follows the semantic proxy, which captures cultural similarity based on the meaning of item descriptions. Second, geographical-based pairing follows the demographic proxy, which reflects cultural closeness based on how near the countries are to each other.

For each item x_i, we find x_j such that it forms two types of pairs: hard pair $\mathbf{p}_{i,j}^{\text{hard}}$, representing the most semantically similar or geographically closest item pair, and the easy pair $\mathbf{p}_{i,j}^{\text{easy}}$, representing the least semantically similar or geographically farthest item pair. Both pair's items always come from the same category.

Description. For description-based pairing, we measure semantic similarity between item descriptions. Each description is encoded into an embedding using the mE5 model [24], and the similarity score between two items is computed using cosine similarity:

$$S(x_i, x_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}.$$

The hardest and easiest pairs for a specific item x_i are defined by finding the index j that minimizes or maximizes this score, respectively, resulting in the pairs $\mathbf{p}_{i,j}^{\text{hard}}$ and $\mathbf{p}_{i,j}^{\text{easy}}$.

$$\begin{aligned} \mathbf{p}_{i,j}^{\text{hard}} &= (x_i, x_j) \quad : \quad j = \arg\min_{k \neq i} S(x_i, x_k), \\ \mathbf{p}_{i,j}^{\text{easy}} &= (x_i, x_j) \quad : \quad j = \arg\max_{k \neq i} S(x_i, x_k). \end{aligned} \tag{1}$$

A higher similarity value indicates that the items share strong semantic overlap, making $\mathbf{p}_{i,j}^{\text{hard}}$ more likely to confuse the model. In this setting, an item x_j serves as an adversarial example for x_i when it closely resembles x_i despite originating from a different culture. For example,

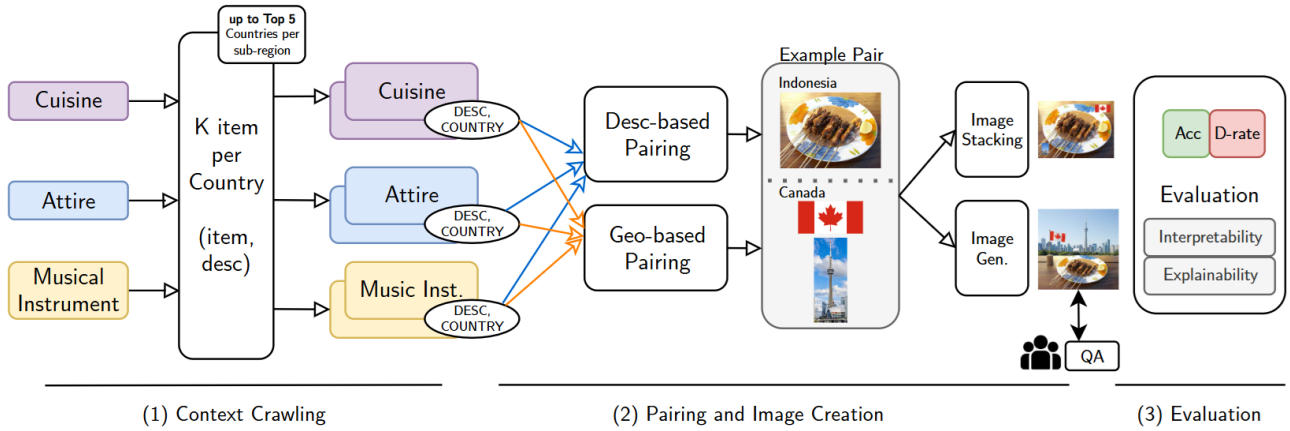


Figure 2. Our CONFUSEDTOURIST construction pipeline. The pipeline consists of 3 stages: **(1) Context Crawling** to obtain balanced, culturally diverse item data and descriptions; **(2) Pair & image creation** where we generate hard and easy cultural pairings and produce various perturbation-infused visual cases; and **(3) Evaluation**, where we assess VLMs' concept grounding ability using objective metrics and interpretability analysis.

lemper from Indonesia and *sushi* from Japan are both rice-based dishes often wrapped in natural leaves or seaweed. Although they belong to distinct culinary traditions, their textual and visual descriptions are closely aligned, making them more likely to confuse the model.

Geographical Distance. For geo-based pairing, we use geographic proximity between countries of origin. Each item x_i is assigned the centroid coordinates $(\text{long}_i, \text{lat}_i)$ of its corresponding country, obtained using the `geopy` library. We denote the distance between two items x_i and x_k as $D(x_i, x_k)$, which is computed using the Haversine formula.

$$\begin{aligned} \mathbf{p}_{i,j}^{\text{hard}} &= (x_i, x_j) : j = \arg\min_{k \neq i} D(x_i, x_k), \\ \mathbf{p}_{i,j}^{\text{easy}} &= (x_i, x_j) : j = \arg\max_{k \neq i} D(x_i, x_k). \end{aligned} \tag{2}$$

The hard pair is expected to consist of items from geographically proximate regions that are more likely to share stylistic or historical influences despite national boundaries. For instance, *batik* from Indonesia and *songket* from Malaysia are produced in neighboring regions and exhibit overlapping textile motifs and weaving traditions. Such pairs may be challenging, making the model struggle to distinguish closely related regional cultures.

3.3. Image Perturbation

We perturb the images using 2 approaches. First, we use an image-generation model, similar to the methodology used in prior work [10]. Second, we employ a simple image stacking perturbation to determine whether a simpler perturbation could already alter the model's accuracy. A handful of visual results produced via both approaches are provided in Appendix I.

3.3.1. Image Stacking Perturbation

We apply this perturbation by stacking the smaller adversarial image over the original cultural item image, which serves as a baseline adversarial attempt to assess models' robustness. We denote adversarial cue images as flag image ($\mathbf{I}_f$) and landmark image ($\mathbf{I}_l$). The size of the adversarial images is resized, denoted by the resizing operation ∇, to maintain their original aspect ratio while ensuring neither dimension exceeds 20% of the original item image dimension. Spatial placement is fixed and non-overlapping: $\mathbf{I}_f$ is consistently placed in the top-right corner ($\nearrow$), and $\mathbf{I}_l$ in the bottom-left corner ($\swarrow$), with a small offset from the edges. This stacking perturbation process, Φ, is defined as a function of the original image $\mathbf{I}_{ori}$ and a set containing either or both of $\mathbf{I}_f$ and $\mathbf{I}_l$:

$$\mathbf{I}_S = \Phi(\mathbf{I}_{ori}, \{\nabla(\mathbf{I}_f), \nabla(\mathbf{I}_l)\}) \tag{3}$$

3.3.2. Generative Perturbation

To curate an alternate perturbation case that is more immersive and naturally integrated, we use an image-generation model, specifically Gemini-2.5-Flash-Image [6]. We denote this process as Ψ resulting in a perturbed image ($\mathbf{I}_G$) as a function of the original item image ($\mathbf{I}_{ori}$), the adversarial cue images ($\mathbf{I}_f, \mathbf{I}_l$), a set containing either or both of $\mathbf{I}_f$ and $\mathbf{I}_l$, and a guiding prompt ($\mathbf{p}$):

$$\mathbf{I}_G = \Psi(\mathbf{I}_{ori}, \{\mathbf{I}_f, \mathbf{I}_l\}, \mathbf{p}). \tag{4}$$

The strict prompt template ($\mathbf{p}$) and preserved inference hyperparameters are employed to best ensure spatial consis-

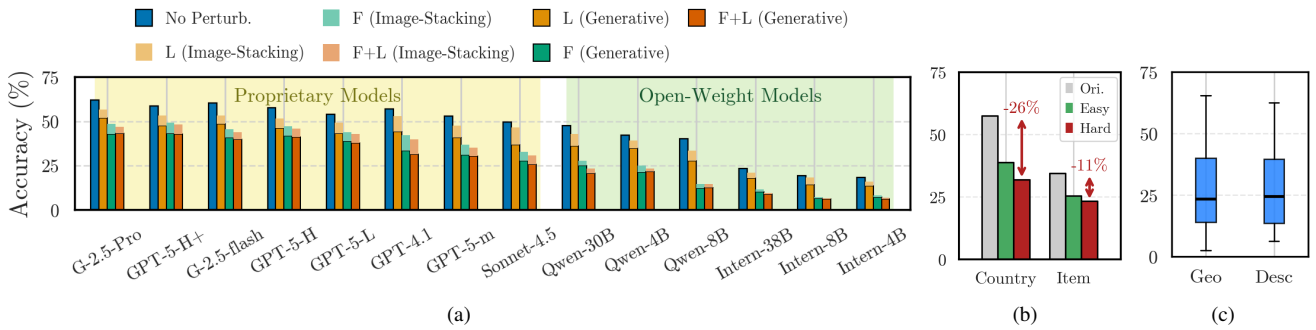


Figure 3. Overall evaluation results in average accuracy of country & cultural item prediction. Key trends: (a) Proprietary VLMs outperform open-weight variants, with generative perturbations being more adverse (especially with flags). (b) Predicting cultural item name is more challenging even from baseline case, though country accuracy drops are much larger in both difficulty levels. (c) Similar average performance of both pairing methods

tency, adversarial semantic guidance, and reproducibility. The details of which are outlined in Appendix C.

In attire-specific cases, our collected images often feature human body parts, posing a potential PII risk. Therefore, before inclusion in Ψ, we apply a preprocessing step where we use an image-generation model to isolate the clothing component. This process removes potential PII and preserves the cultural attire in a more neutral form. This enables subsequent adversarial perturbations without compromising privacy (see Appendix C for details).

We also conduct a manual quality check of the generative perturbed images. A random sample of 5% from each category: cuisine, attire, and musical instrument was selected, as denoted previously with $\mathbf{I}_G$. Using the rubric defined in Appendix D, the authors independently re-evaluated 130 sampled images, assigning scores on a 1–5 Likert scale. The generated images achieved an average score of 4.49, demonstrating the high quality of the perturbed outputs. However, quality varied across categories, with cuisine being the most realistic and attire the least.

3.4. Evaluation Metrics

To evaluate the models' robustness and vulnerability to cultural perturbations, we use two metrics: one that measures general prediction accuracy across all categories, and another that measures the model's distractive behavior on incorrect country predictions.

Multi-Target Accuracy (Acc.). This metric measures substring match accuracy when each instance may have multiple ground-truth labels (e.g., alternative country or cultural item names). A prediction (p_i) is considered correct if the uncased prediction string is a substring of any uncased ground-truth string(s) ($\mathbf{G}_i$). The overall Accuracy (Acc.) is the average number of successful substring matches across all N instances.

$$\text{Acc.} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & \text{if } \exists\, g \in \mathbf{G}_i \text{ s.t. } p_i \subseteq g \\ 0 & \text{otherwise}\end{cases} \tag{5}$$

Model Distraction Likelihood ($D_{\mathcal{L}}$). For each model, this metric measures the frequency with which an incorrect country prediction is directly deceived by the counterpart adversarial cue. This metric is essential for assessing model robustness because it quantifies the causal attribution of the error, rather than just counting general mistakes. For example, if a model is tasked with predicting Country A but is perturbed with content from Country B, a high score on this metric indicates that the model's error is a direct result of being swayed by the adversarial content from Country B, demonstrating a critical failure in visual grounding against geographic conflict.

$$D_{\mathcal{L}} = P(\text{adv} \mid \text{wrong}) \tag{6}$$

4. Evaluation & Results

This section explains our evaluation methodology using ConfusedTourist and presents the main results. We cover the model settings, the strict single-turn prompt protocol used for inference, as well as the core findings.

4.1. Experiment Setup

Model Selection. We employ 14 model settings to evaluate current state-of-the-art VLMs. These settings are comprised of 8 proprietary variants, which spans across 3 model families (GPT, Gemini, Claude). We employ the other 6 open-weight settings that are part of 2 open-weight families, including Qwen3VL [16] and InternVL3.5 [25].

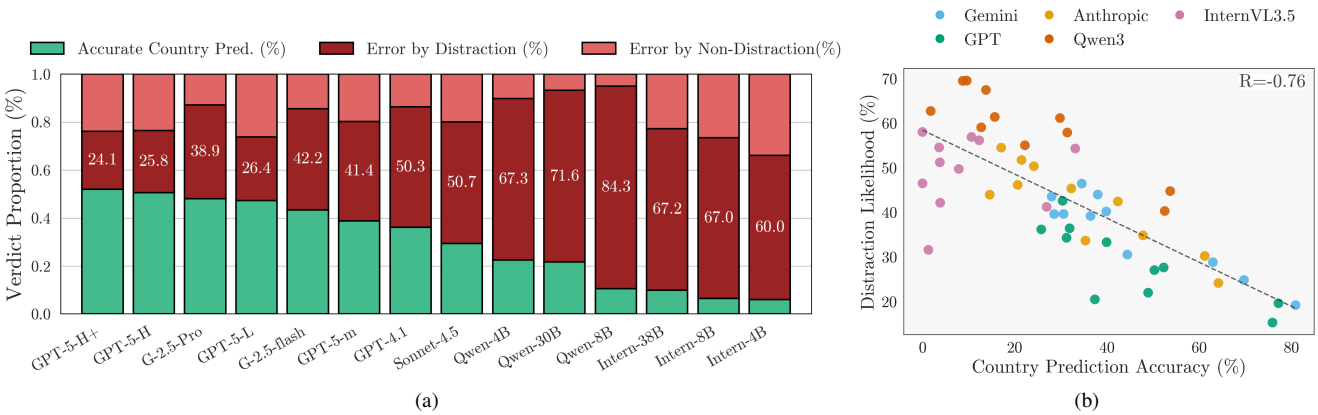


Figure 4. The negative correlation between country prediction accuracy vs. distraction likelihood of the model in wrongly predicted cases. (a) The proportion of wrongly predicted countries across models increases along with the decrease of country prediction accuracy. (b) Across 11 different subregions for each VLM, the correlation of this relationship is also scoring at -0.76, suggesting a strong negative relation between the metrics.

Prompt. All instances are being evaluated via a single inference call. VLMs are instructed to provide the answer for both the *name* of the target cultural item and its *country of origin*, as shown in the prompt below, where {`category`} is a choice between `attire`, `cuisine`, or `musical instrument`. We provide the detailed inference hyperparameters for each model in Appendix G.

```
Observe the image and determine the original name of
this {category} object, and the country from which this
{category} originally comes. Return the original name
of this {category} object first, followed by the country
name.
```

4.2. Evaluation Result

In this section, we outline the most apparent findings across all of our features. While the complete evaluation results can be observed in Table 4 and 5, we summarize the overall trend in major features in the points below by referring to Figure 3:

(1) **Superior performance of proprietary VLMs**. Proprietary models outperform their open-weight variants, with the open-weight Qwen family scoring the closest to them, as depicted in Figure 3a. The difference in average prediction accuracy across the perturbation contexts (F, L, F+L) between the best performing GPT and Qwen3-VL variants is 22% for country prediction and 27% for item prediction. There is no significant benefit from the model's reasoning mode in either the baseline or in showing any improved resistance against adversarial context.

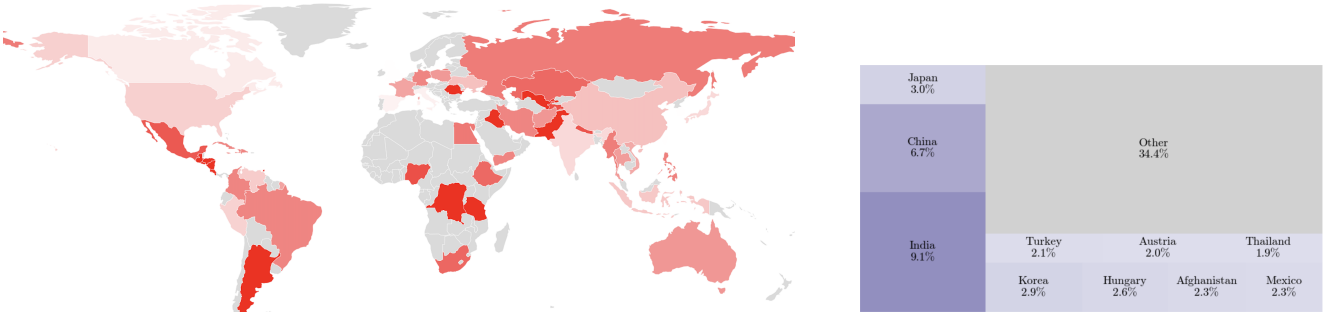
(2) **Flag object is the main perturbation driver.** In the image stacking setting, the presence of the flag caused a decline of up to 18.4%, whereas applying the landmark perturbation resulted in a minor drop of only up to 6.9%. Furthermore, the combination of their perturbation effect is observed to be no more than the sum of the individual parts, indicating that no emergent adversarial combination resulted from such combination.

(3) **Generative perturbation is more effective**, leading to an average performance worsening of 17.34%, compared to only 8.43% drop using image-stacking method. This validates that the generative (versus image-stacking) perturbation method is more successful in tricking the model. As observed in Figure 3a, the numbers are consistent in all cases when comparing shades within the same color in each perturbation context bar (F, L, F+L).

(4) **Proxy-based semantic and geographic cultural proxies remain relevant**. As observed in Figure 3b, the difference in average accuracy drop is up to 26% in the hardest cases (country) across cases. This signals that distinguishing morphologically similar items and items from geographically close cultures remains a challenge due to concept overlaps. Aligning with prior cultural proxy selections from Adilazuarda et al. [1], our pairing method also (description versus geographic-based) yields consistent results, as shown in Figure 3c. This further suggests that both semantic and geographic cultural proxies remain relevant in multimodal multicultural evaluation.

5. Discussion & Analysis

In this section, we outline more fine-grained observations, where we go deeper discussing nature of observed visual distraction, existence of systematic country fallback biases, and the thought process of the VLMs in doing predictions.



(a) (b)

Figure 5. GPT-5 (High$^+$) results. (a) Global map of accuracy drop, computed as the ratio between performance difference and original score. (b) Distribution of predicted countries where the model is incorrect but does not follow the adversarial country

Dominant Effect of Adversarial Cues in Prediction Errors. We observe a clear inverse relationship: the lower a model's country prediction accuracy, the greater the tendency for its inaccurate predictions to drift directly toward our perturbation cues. Figure 4a depicts this trend, highlighting that the decline in general accuracy is proportionate to the specific inaccuracy caused by distractive cues. This is further amplified by the significant negative correlation ($R = -0.76$) measured between these features across all cases, as shown in the scatter plot in Figure 4b.

Amidst these findings, the Qwen family presents a compelling insight: despite maintaining a higher overall accuracy compared to InternVL3.5, it exhibits a higher Distraction Likelihood ($D_{\mathcal{L}}$). This finding suggests a pronounced their vulnerability to the adversarial cue specifically in error scenarios. We leverage this focused behavior—that their errors are less random—to select its 30B variant for deeper interpretability analysis, as it may help reducing the noise in the attention spread visualization.

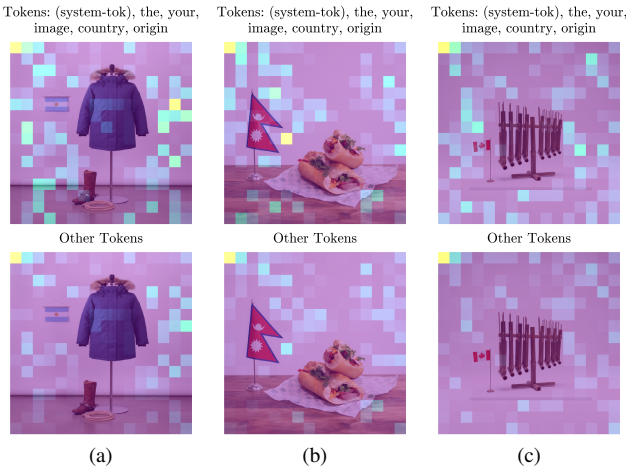
Our interpretability analysis revealed a critical finding: the model's attention focus on the image tokens was disproportionately driven by specific text components—namely, system tokens, geo-related tokens, and a subset of stopwords (as shown in Figure 6). We further conducted an ablation study to investigate the effect of these tokens as detailed in Appendix H. Critically, eliminating these suspected tokens resulted in two key improvements: a measurable increase in grounding accuracy and a pronounced shift in attention back to the intended image region. While this suggests an avenue for better prompt choice selection for our work, we still conclude that:

1. VLMs tend to rely on easily interpretable visual cues (e.g., flags or other prominent context features), leading to failures in incorporating relevant knowledge or being overridden by familiar visual shortcuts.
2. This behavior worsens due to reliance on specific text tokens, showing that even advanced VLMs can be unstable and highly sensitive to prompts—as seen in one failure case (Appendix H, attire), where a prompt correction even turned a previously correct answer into a wrong one.

These findings express the need for more globally aware models that remain stable under small, semantic-based input changes (i.e. in our case, specific geographic cues).

Specific Country Fallback Bias. As shown in Figure 5(a), GPT-5 (High$^+$) exhibits larger performance drops in several low-resource regions such as Africa, Latin America, and MENA. In contrast, supposedly low-resource regions like Southeast Asia (SEA) show relatively stable performance, possibly reflecting improved data coverage from recent community efforts such as SEACrowd [13] and SEA-VL [4].

To further investigate regional prediction tendencies, we



(a) (b) (c)

Figure 6. Attention heatmap analysis indicates that visual grounding primarily arises from a limited set of tokens. In (a) attire, (b) cuisine, and (c) musical-instrument culture items, tokens linked to system cues, geographic references, and category-specific terms dominate the model's visual attention.

analyze cases where the model's predictions are incorrect but do not correspond to the adversarial country $\mathbf{C}_{adv}$. This scenario corresponds to the negation of $D_{\mathcal{L}}$ (Eq. 6). As shown in Figure 5b, certain countries are predicted more frequently even in the absence of adversarial cues, suggesting a "fallback preference" toward specific regions.

For example, GPT-5 (High^{+}) tends to over-predict countries such as India (9.1%), China (6.7%), and Japan (3.0%). Notably, India and China also appear in the top three most frequent incorrect predictions for all other models evaluated. This consistency suggests that image-text pairs from these cultures are more prevalent in the training distribution. We hypothesize that this over-representation may arise because these regions are common subjects of online imagery (e.g., as popular tourist destinations).

Flag Cues Corrupt Model Reasoning Traces. We also analyze the model outputs by tracing the reasoning process of the Gemini model in sample cases where it produces incorrect answers. Our comparison considers several factors: perturbation type (flag, landmark, flag+landmark), category (cuisine, attire, musical instrument), and perturbation method (image-stacking vs. generated perturbation). From our observations, the model often exhibits uncertainty in the middle of its reasoning steps, with phrases resembling "Am I correct? I am not sure".

Flag perturbation — Reasoning
"The style of this garment screams West or Central African influence to me. The wide sleeves, the embroidery... I'm leaning heavily towards the DRC (from the flag). I need to find the specific name for this type of clothing in the DRC."

Landmark perturbation — Reasoning
"I'm thinking thobe, dishdasha, kandura, maybe even jalabiya. Could even be a simplified caftan, given the Turkish context. Since the prompt is about the origin (not where it's displayed), I need to be more precise."

Flag + Landmark perturbation — Reasoning
"I see a white blouse with red and black embroidery on a mannequin. There's a flag visible — a vertical tricolor of blue, yellow, and red (Romania). The background has that massive building — the Palace of the Parliament (Bucharest). Definitely Romania."

Among these factors, the most distinct differences appear in the perturbation type. The model typically focuses on the flag first before attending to other relevant aspects of the image. Landmark perturbations are generally less effective, as the model sometimes concludes that "there is no connection between the landmark and the object", leading it to refocus on the primary object instead. However, in flag+landmark cases, the model tends to perceive a strong correlation between the flag and the landmark, and thus shifts its reasoning toward these contextual cues rather than the main object itself.

In contrast, the category and perturbation method do not appear to introduce substantial differences, as the model's reasoning predominantly centers around the flag across most cases.

6. Conclusion

We introduced CONFUSEDTOURIST , a novel cultural adversarial robustness suite designed to evaluate VLMs' ability to accurately identify cultural items in adversarially-induced images. Our experiments reveal a critical vulnerability: all state-of-the-art VLMs experience a substantial accuracy drop under simple image stacking, which becomes even more severe under generative perturbations. We found these geographical-induced perturbations consistently cause disruption across all cases, a vulnerability that is more prominent with the presence of a flag, which consistently exhibits the model's grounding bias toward the adversarial cue. Further analysis shows that as model accuracy decreases, the models become increasingly distracted by these perturbations, focusing more on the distractor cues than on the cultural item itself, with the reasoning traces supporting this observation. Overall, this work highlights a critical challenge: VLMs must develop greater cultural robustness to achieve reliable multimodal understanding across diverse cultural contexts.

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3, 6

[2] Ashwath Vaithinathan Aravindan, Abha Jha, and Mihir Kulkarni. Do vlms have bad eyes? diagnosing compositional failures via mechanistic interpretability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 704–712, 2025. 2

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. 2

[4] Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib,

Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, Mohamed Fazli Mohamed Imam, Priyaranjan Pattnayak, Salsabila Zahirah Pranida, Kevin Pratama, Yeshil Bangera, Adisai Na-Thalang, Patricia Nicole Monderin, Yueqi Song, Christian Simon, Lynnette Hui Xian Ng, Richardy Lobo Sapan, Taki Hasan Rafi, Bin Wang, Supryadi, Kanyakorn Veerakanjana, Piyalitt Ittichaiwong, Matthew Theodore Roque, Karissa Vincentio, Takdanai Kreangphet, Phakphum Artkaew, Kadek Hendrawan Palgunadi, Yanzhi Yu, Rochana Prih Hastuti, William Nixon, Mithil Bangera, Adrian Xuan Wei Lim, Aye Hninn Khine, Hanif Muhammad Zhafran, Teddy Ferdinan, Audra Aurora Izzani, Ayushman Singh, Evan Evan, Jauza Akbar Krito, Michael Anugraha, Fenal Ashokbhai Ilasariya, Haochen Li, John Amadeo Daniswara, Filbert Aurelian Tjiaranata, Eryawan Presma Yulianrifat, Can Udomcharoenchaikit, Fadil Risdian Ansori, Mahardika Krisna Ihsani, Giang Nguyen, Anab Maulana Barik, Dan John Velasco, Rifo Ahmad Genadi, Saptarshi Saha, Chengwei Wei, Isaiah Edri W. Flores, Kenneth Chen Ko Han, Anjela Gail D. Santos, Wan Shen Lim, Kaung Si Phyo, Tim Santos, Meisyarah Dwiastuti, Jiayun Luo, Jan Christian Blaise Cruz, Ming Shan Hee, Ikhlasul Akmal Hanif, M.Alif Al Hakim, Muhammad Rizky Sya'ban, Kun Kerdthaisong, Lester James Validad Miranda, Fajri Koto, Tirana Noor Fatyanosa, Alham Fikri Aji, Jostin Jerico Rosal, Jun Kevin, Robert Wijaya, Onno P. Kampman, Ruochen Zhang, Börje F. Karlsson, and Peerat Limkonchotiwat. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria, 2025. Association for Computational Linguistics. 7

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1

[6] Google. Gemini 2.5 Flash Image Model (Nano Banana). https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. Informally known as Nano Banana; Accessed on: [Insert Access Date]. 4

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. 2

[8] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 2

[9] Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2

[10] Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. WHEN TOM EATS KIMCHI: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2, 3, 4

[11] Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture, 2024. 2

[12] Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*, 2025. 2

[13] Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7

[14] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024. 2

[15] Beth Pearson, Bilal Boulbarss, Michael Wray, and Martha

Lewis. Evaluating compositional generalisation in vlms and diffusion models, 2025. 2

[16] Alibaba Cloud Qwen Team. Qwen3-vl. https://github.com/QwenLM/Qwen3-VL, 2025. Accessed: 2025-11-11. 5

[17] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024. 2, 3

[18] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. 2

[19] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024. 2, 3, 1

[20] Burak Satar, Zhixin Ma, Patrick A. Irawan, Wilfried A. Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. Seeing culture: A benchmark for visual reasoning and grounding, 2025. 2, 3

[21] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Amirudin, Muhammad Ridzuan, Daniya Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Khan. All languages matter: Evaluating lmms on culturally diverse 100 languages, 2024. 2

[22] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Hafizi Amirudin, Muhammad Ridzuan, Daniya Najiha Abdul Kareem, Ketan Pravin More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Ballah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Augusto Zacarias Xavier, Amit Bhatkal, Hawau Olamide Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Shahbaz Khan. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575, 2025. 2

[23] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025. 2, 3

[24] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024. 3

[25] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. 5

[26] Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Christabelle Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie

Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2, 3, 1

[27] Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. CLAIM: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13080–13094, Vienna, Austria, 2025. Association for Computational Linguistics. 2, 3

[28] Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. Cultiverse: Towards cross-cultural understanding for paintings with large language model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6710–6719, 2025. 2