

Seeing Symbols, Missing Cultures: Probing Vision-Language Models’ Reasoning on Fire Imagery and Cultural Meaning

Haorui Yu

Duncan of Jordanstone College of
Art & Design (DJCAD), University of Dundee
Dundee, United Kingdom
2655435@dundee.ac.uk

Yijia Chu

Faculty of Arts, Xiamen University
Xiamen, China
18620221154827@stu.xmu.edu.cn

Yang Zhao

Guangzhou Institute of Science
and Technology (GZIST), Guangzhou, China
zhaoyang@gzist.edu.cn

Qiufeng Yi

University of Birmingham
Birmingham, United Kingdom
qxy953@student.bham.ac.uk

Abstract

Vision-Language Models (VLMs) often appear culturally competent but rely on superficial pattern matching rather than genuine cultural understanding. We introduce a **controlled** diagnostic framework to probe VLM reasoning on fire-themed cultural imagery through both classification and explanation analysis. Testing multiple models on Western festivals, non-Western traditions, and emergency scenes reveals systematic biases: models correctly identify prominent Western festivals but struggle with underrepresented cultural events, frequently offering vague labels or misclassifying emergencies as celebrations. These failures **pose risks in public-facing or safety-critical applications** and highlight the need for **explanation-driven** cultural evaluation beyond accuracy metrics to support interpretable and fair multimodal systems.

1 Introduction

Vision-Language Models (VLMs) demonstrate sophisticated capabilities, often appearing culturally aware by correctly identifying festivals and artifacts (Sukienik et al., 2025; Liu et al., 2025). This apparent competence creates a “semantic illusion” where pattern matching masquerades as understanding (Li et al., 2023). A model might label an image as “Torch Festival” not from understanding its Yi ethnic significance, but from associating visual cues like fire and crowds with festival tokens.

This surface-level pattern matching creates critical vulnerabilities (Ananthram et al., 2024).

Common visual elements are semantically ambiguous and culturally polysemous (Saussure, 1916; Turner, 1967). Fire can signify celebration, crisis, or ritual transformation across cultures (Bachelard, 1964). When VLMs use “symbolic shortcuts”—defaulting to familiar associations rather than contextual specificity—they risk misinterpreting cultural meaning (Blodgett et al., 2020). Models unable to distinguish Peru’s sacred Inti Raymi from Britain’s Lewes Bonfire, or from dangerous fires, **pose risks in public-facing or safety-critical applications and therefore warrant additional cultural-robustness evaluation before deployment** (Mehrabi et al., 2021).

This paper investigates whether VLMs understand cultural semantics or rely on symbolic shortcuts. We extend recent work on VLM cultural biases (Nayak et al., 2024; Qiu et al., 2025) with a diagnostic framework probing reasoning patterns. We analyze both classification labels and explanations (Ferrara, 2024), evaluating models on visually similar but semantically distinct fire-themed images to expose reasoning failures that accuracy metrics miss. Figure ?? illustrates our methodology.

Our approach differs from recent frameworks like CROSS (Qiu et al., 2025) in three key ways: (1) we focus on explanation analysis rather than accuracy alone, (2) we **use “symbolic shortcuts”**

as our diagnostic lens (rather than claiming a new concept), and (3) we identify safety-critical failure modes when cultural misinterpretation occurs in emergency contexts.

We formally define **symbolic shortcuts** as reasoning patterns where models map visual elements (e.g., fire) directly to their most common semantic associations (e.g., festival) while neglecting contextual cues that would enable proper cultural interpretation. **Rather than a comprehensive benchmark, our contribution is a controlled diagnostic focused on a single multi-meaning symbol (“fire”), complementary to breadth-first cultural evaluations.**

Our key contributions are: (1) A diagnostic framework that moves beyond accuracy to evaluate VLM reasoning through classification and explanation. (2) A targeted analysis revealing how symbolic shortcuts lead to cultural misinterpretations and safety risks. (3) Evidence of a significant reasoning gap between Western and non-Western cultural contexts, highlighting data bias and fairness issues in state-of-the-art models.

2 Symbolic Reasoning Probe

Our diagnostic framework is designed to probe the reasoning behind a VLM’s cultural classifications. It assesses whether a model’s output is based on genuine semantic understanding or a reliance on superficial visual cues. The probe consists of three components: a curated dataset with controlled semantic ambiguity, a selection of diverse VLMs, and an evaluation protocol that demands both classification and explanation.

2.1 Model Selection

We evaluate **9** recent Vision-Language Models (5 proprietary and 4 open-source), representing a diverse range of architectures and developers. This selection allows for a comprehensive comparison across the most capable models available at the time of study.

2.2 Dataset

To test the models’ ability to handle symbolic ambiguity, we curated a Multi-Cultural Heritage Dataset (MCHD) of 77 images. The images are thematically consistent (fire-related) but semantically diverse, organized into three categories designed to challenge superficial visual reasoning:

- **Modern Western Festivals (e.g., Burning**

Model	Type	Developer
GPT-4o	Proprietary	OpenAI
Claude 3.5 Haiku	Proprietary	Anthropic
Claude 3.7 Sonnet	Proprietary	Anthropic
Claude 4 Sonnet	Proprietary	Anthropic
Claude 4 Opus	Proprietary	Anthropic
Aya Vision 32B	Open-source	Cohere
Aya Vision 8B	Open-source	Cohere
Qwen2.5-VL 72B	Open-source	Alibaba
Qwen2.5-VL 7B	Open-source	Alibaba

Table 1: Vision–Language Models evaluated (**9** total; 5 proprietary, 4 open-source).

Man, Guy Fawkes Night): Events with extensive documentation and high representation in typical training data.

- **Underrepresented Non-Western Traditions (e.g., Huobajie, Sadeh, Inti Raymi):** Events that are visually similar to Western festivals but have distinct cultural meanings and are less likely to be well-represented in training corpora.
- **Non-Cultural Emergencies (e.g., wildfires, structural fires):** Scenes containing fire and sometimes crowds, serving as a critical control group to test for cultural misattribution and safety-critical failures.

A detailed list of the cultural traditions included is available in Appendix A.1.

The 77 images were sourced from publicly available online repositories under Creative Commons licenses. Selection criteria included: (1) clear fire-related visual elements, (2) sufficient contextual cues for cultural identification, (3) resolution suitable for VLM processing (minimum 512×512 pixels), and (4) verification of cultural authenticity through multiple sources. The distribution comprises: 30 Western festival images, 37 non-Western tradition images, and 10 emergency control images.

Availability. To prevent test contamination and overfitting in future model training, we keep the test images private while releasing metadata (URLs, licenses, cultural labels) and the full evaluation scripts/prompts in the supplementary material.

2.3 Evaluation Protocol

We use a single, zero-shot direct prompt for both classification and explanation: “Please identify the cultural event or tradition shown in this image. Provide a specific name and general category.”

This simple prompt **aims to minimize prompt-engineering confounds and tests inherent reasoning without tailored instructions**. We acknowledge that our prompt “Please identify the cultural event or tradition” may introduce bias by priming models toward cultural interpretations. Future work will explore more neutral prompts such as “What is shown in this image?” to reduce presuppositional influences on model responses.

We then manually analyzed the textual responses following these criteria:

- **Symbolic shortcuts:** When models rely on generic visual features (e.g., “fire equals festival”)
- **Cultural-specific knowledge:** When explanations include specific cultural details

Analysis was conducted independently by two evaluators with inter-rater reliability of 0.87 (Cohen’s kappa), with disagreements resolved through discussion.

Design rationale (stress test). We intentionally use a presuppositional single-step prompt as a *worst-case stress test* that mirrors common user flows and probes whether models can resist cultural priming when the image is in fact an emergency. Without changing prompts, we also report GPT-4o results for two releases (08–06 and 11–20) alongside their aggregate; the qualitative and quantitative patterns are consistent across versions (Tables 2, 6, 7).

Due to the subjective nature of cultural recognition and the specialized knowledge required, establishing human baselines is beyond the scope of this diagnostic study.

3 Findings

State-of-the-art VLMs consistently favor symbolic shortcuts over cultural reasoning, evident in qualitative output differences and varied performance across categories.

Table 3 contrasts GPT-4o and Qwen2.5-VL 72B outputs. Both correctly identify Burning Man, but for Huobajie, GPT-4o identifies the Yi ethnic tradition while Qwen provides only “Bonfire festival”—demonstrating cultural knowledge gaps.

Three primary failure modes emerge: (1) **Cultural Misclassification**—labeling emergencies as cultural events; (2) **Generic Labeling**—using

vague descriptors; (3) **Western-centric Bias**—defaulting to familiar Western events.

Critically, Qwen misinterprets a wildfire as Guy Fawkes Night—a safety-critical failure where models hallucinate familiar cultural contexts onto dangerous events (see Appendix A.3).

Tables 2 and 3 quantify this imbalance and demonstrate the qualitative differences. Performance on `burning_man_amERICAN` reaches 100%, but drops to 0% for `sadeh_iranian` and `huobajie`, revealing bias toward Western, internet-prominent events.

Figure 1 visualizes failure patterns. GPT-4o shows distributed errors, while Qwen2.5-VL 7B systematically defaults to `guy_fawkes` or `burning_man` when uncertain—*consistent with* reliance on symbolic shortcuts.

4 Discussion

Our findings reveal a gap between visual pattern recognition and cultural understanding. Models’ “symbolic shortcuts”—overgeneralizing fire as festival—create competence illusions masking reasoning failures.

Data imbalance drives these failures. Superior performance on Burning Man versus poor results on Huobajie and Sadeh reveals Western-centric training data bias (Ferrara, 2024). Models learn simplified dominant representations, not varied cultural meanings.

Mechanistic hypothesis (post hoc). The confusion patterns (Fig. 1) suggest that, under uncertainty, some models disproportionately map inputs to frequent Western tokens (e.g., `guy_fawkes`, `burning_man`), a “shortcut prior” in which co-occurring proxies (flames, crowd density, nighttime) outweigh contextual cues. This is consistent with spurious-correlation phenomena discussed in fairness/bias surveys (Ferrara, 2024; Mehrabi et al., 2021; Blodgett et al., 2020). A full mechanistic dissection (e.g., attribution analyses) is beyond our diagnostic scope and left for future work.

This causes cultural erasure—labeling Celtic Samhuinn as “bonfire”—and safety failures—misclassifying wildfires as festivals. Systems unable to distinguish celebration from catastrophe pose risks in public-facing or safety-critical applications and therefore warrant additional cultural-robustness evaluation before deployment.

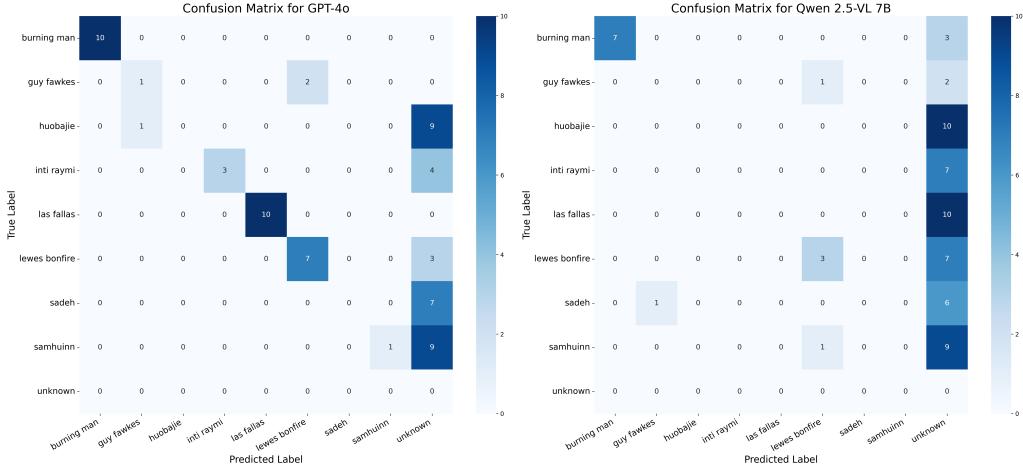


Figure 1: Confusion Matrices for GPT-4o (left) and Qwen 2.5-VL 7B (right). The rows represent the true cultural labels, and the columns represent the predicted labels. These matrices reveal the specific patterns of misclassification for the highest and lowest-performing models.

Tradition	GPT-4o	Claude	Qwen*
Burning Man	100.0	80.0	80.0
Guy Fawkes	33.3	66.7	0.0
Huobajie	0.0	0.0	0.0
Inti Raymi	42.9	28.6	14.3
Las Fallas	100.0	100.0	50.0
Lewes Bonfire	70.0	50.0	20.0
Sadeh	0.0	14.3	0.0
Samhuinn	10.0	0.0	10.0

Table 2: Fine-grained accuracy by cultural category (%). Top-tier proprietary (GPT-4o), mid-tier proprietary (Claude 3.7), and open-source (Qwen2.5-VL) models. Best per category in bold. *72B version.

Case Type	Image	GPT-4o (Top-tier)	Qwen2.5-VL 72B
Western		Name: Burning Man Category: Art/music Analysis: Correct	Name: Burning Man Category: Music Analysis: Less specific
Non-Western		Name: Huobajie Category: Yi ethnic Analysis: Accurate	Name: Bonfire fest. Category: Traditional Analysis: Generic
Emergency		Name: Wildfire Category: Emergency Analysis: Correct	Name: Guy Fawkes Category: Festival Analysis: Dangerous

Table 3: Qualitative case study comparing model outputs on three image types, showing disparities in specificity, cultural knowledge, and safety-critical distinctions. **Ground Truth:** Western: Burning Man; Non-Western: Yi Torch Festival (Huobajie); Emergency: **Uncontrolled large-scale outdoor fire (non-cultural)**.

We must shift from accuracy to interpretability, probing *why* models conclude. Scaling current approaches reinforces biases; future work needs cultural context modeling and reasoning-focused evaluation.

4.1 Future Directions

Future research should explore: (1) extending this framework to other cultural domains (clothing, architecture, cuisine) to validate generalizability, (2) developing training methods to mitigate symbolic shortcuts through culture-aware data augmentation, and (3) integrating cultural knowledge graphs into VLM architectures for enhanced contextual reasoning.

5 Conclusion

This paper introduced a diagnostic probe to move beyond accuracy-based evaluation and assess the cultural reasoning of Vision-Language Models. Our findings reveal that current models, including state-of-the-art systems like GPT-4o, often rely on “symbolic shortcuts,” leading to a superficial understanding that fails in nuanced, non-Western, or safety-critical contexts. They can see the symbols, but they often miss the culture.

We argue for a crucial transition in how we evaluate AI systems for cultural tasks: a shift from measuring *what* they classify to understanding *how* and *why* they reason. This explanation-driven approach is essential for identifying fairness risks

associated with data bias and for building models that are not only accurate but also genuinely and safely culturally aware. This work provides a framework and a baseline for this necessary next step in AI development.

Limitations

Our narrow focus on fire festivals ensures consistency but limits generalization to other cultural domains. The 77-image sample, while sufficient to demonstrate our diagnostic framework’s validity, constrains the universality of our conclusions. This work should be viewed as a proof-of-concept for a diagnostic tool rather than a comprehensive evaluation of VLM cultural understanding.

The single-prompt evaluation approach, though revealing, presents opportunities for expansion with varied prompting strategies. Future work could explore prompt variations to assess their impact on cultural recognition. Additionally, broader cultural domains (clothing, architecture, cuisine) and larger datasets would strengthen the generalizability of our findings.

Acknowledgments

We acknowledge the cultural communities whose traditions form this research foundation and thank cultural consultants for their expertise. We appreciate WiNLP’s inclusive research environment.

References

- Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.
- Gaston Bachelard. 1964. *The Psychoanalysis of Fire*. Beacon Press, Boston.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Emilio Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics.
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.
- Ninareh Mehrabi, Fred Morstatter, and Nripsuta et al. Saxena. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.
- Shravan Nayak, Haotian Liu, and Jiasen et al. Lu. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.
- Haoyi Qiu, Kung-Hsiang Huang, Ruichen Zheng, Jiao Sun, and Nanyun Peng. 2025. Cross: Cultural safety evaluation for vision-language models. *arXiv preprint arXiv:2505.14972*.
- Ferdinand de Saussure. 1916. *Course in General Linguistics*. Columbia University Press, New York.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*.
- Victor Turner. 1967. *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca.

A Appendix: Supplementary Materials

A.1 Cultural Categories in MCHD

The Multi-Cultural Heritage Dataset (MCHD) used in this study includes images from the following cultural traditions and control category.

- **Western Traditions:** Burning Man (American), Guy Fawkes Night (British), Las Fallas (Spanish), Lewes Bonfire (English), Samhuin (Scottish).
- **Non-Western Traditions:** Huobajie (Chinese), Inti Raymi (Peruvian), Sadeh (Iranian).
- **Control Group:** Fire Emergencies (e.g., wild-fires, structural fires).

A.2 Qualitative Comparison of Model Explanations

Table 4 provides representative model explanations demonstrating the contrast between symbolic shortcuts and cultural understanding.

A.3 Systematic Cultural Misclassification of Emergencies

Critical safety failures occurred when models misinterpreted emergency scenes as cultural events. The following table documents instances where VLMs classified fire emergencies as festivals, highlighting the danger of symbolic shortcuts in safety-critical applications.

A.4 Comprehensive Fine-Grained Accuracy Benchmark

Full performance metrics across all 9 and 8 cultural traditions plus control group are presented below. These results demonstrate systematic biases toward Western, internet-prominent events.

Image Type	Model	Full Model Explanation (Illustrative)
Non-Western Tradition (Huobajie)	GPT-4o	Prediction: Huobajie (Torch Festival), Folk festival of the Yi people. Reasoning: The image displays elements consistent with the Torch Festival, including large bonfires, traditional clothing worn by participants that resembles Yi ethnic attire, and a celebratory nighttime atmosphere unique to this cultural event.
	Qwen2.5-VL 72B	Prediction: Bonfire festival, Traditional festival. Reasoning: This image shows a large bonfire at night with many people gathered around. These are typical features of a bonfire festival.
Non-Cultural Emergency (Wildfire)	GPT-4o	Prediction: Forest fire / Wildfire, Emergency event. Reasoning: The image depicts an uncontrolled fire spreading through a forest. This is a characteristic scene of a wildfire, which is a natural disaster, not a cultural event.
	Qwen2.5-VL 72B	Prediction: Guy Fawkes Night, Festival. Reasoning: The large fire in the image is reminiscent of the bonfires traditionally lit during Guy Fawkes Night celebrations in the UK. The event appears to be a public gathering.

Table 4: Qualitative comparison of full textual explanations generated by a top-tier proprietary model (GPT-4o) and a leading open-source model (Qwen2.5-VL 72B). The examples illustrate GPT-4o’s ability to cite specific cultural knowledge versus Qwen’s reliance on generic visual cues, which leads to critical misclassification of an emergency.

Model	Error Type	Prediction	Impact
<i>Cultural Misclassification (Safety-Critical)</i>			
Claude 3.7 Sonnet	Emergency	Las Fallas	Misinterprets danger as celebration
Aya Vision 8B	Emergency	Guy Fawkes	Could delay emergency response
Aya Vision 32B	Emergency	Guy Fawkes	Could delay emergency response
Qwen2.5-VL 7B	Wildfire	Guy Fawkes	Misses critical safety context
Qwen2.5-VL 72B	Wildfire	Burning Man	Normalizes dangerous situation

Table 5: Instances of Cultural Misclassification where models incorrectly identified non-cultural fire emergencies as cultural festivals. This table highlights the safety-critical implications of these failures, which are particularly prevalent in open-source models.

Proprietary Models	Burning Man	Guy Fawkes	Huobajie	Inti Raymi	Avg.
GPT-4o*	100.0	33.3	0.0	42.9	44.1
Claude 3.7 Sonnet	80.0	66.7	0.0	28.6	43.8
Claude 4 Opus	100.0	66.7	0.0	28.6	48.8
Claude 3.5 Haiku	80.0	66.7	0.0	28.6	43.8
Claude 4 Sonnet	100.0	100.0	0.0	14.3	53.6
Open-Source Models					
Aya Vision 32B	60.0	0.0	0.0	0.0	15.0
Aya Vision 8B	80.0	33.3	0.0	28.6	35.5
Qwen2.5-VL 72B	80.0	0.0	0.0	14.3	23.6
Qwen2.5-VL 7B	80.0	0.0	0.0	0.0	20.0

Table 6: Performance comparison on cultural categories (Part 1). *GPT-4o results averaged across versions.

Model	Las Fallas	Lewes Bonfire	Sadeh	Samhuinn	Emergencies
Proprietary					
GPT-4o*	100.0	70.0	4.8	6.7	100.0
Claude 3.7 Sonnet	100.0	50.0	14.3	0.0	90.0
Claude 4 Opus	100.0	40.0	0.0	0.0	100.0
Claude 3.5 Haiku	100.0	40.0	0.0	0.0	100.0
Claude 4 Sonnet	100.0	20.0	0.0	0.0	90.0
Open-Source					
Aya Vision 32B	25.0	10.0	0.0	0.0	90.0
Aya Vision 8B	0.0	0.0	0.0	0.0	80.0
Qwen2.5-VL 72B	50.0	20.0	0.0	10.0	70.0
Qwen2.5-VL 7B	0.0	30.0	0.0	0.0	80.0

Table 7: Performance comparison on cultural categories (Part 2) and emergency control set.