Lost in Translation: A Position Paper on Probing Cultural Bias in Vision-Language Models via Hanbok VQA

Anonymous ICCV submission

Paper ID 10

Abstract

While Vision-Language Models (VLMs) offer transformative potential for cultural heritage preservation, they often exhibit significant "cultural blind spots" due to training data heavily skewed towards Western contexts. This leads to limited understanding of non-Western cultures, such as those from East Asia. This paper posits that modern VLMs consequently fail to accurately interpret underrepresented cultural objects, leading to misidentification, cultural confusion, and factual hallucination. To investigate this, we evaluate prominent VLMs including LLaVA-1.5, ViP-LLaVA, Shikra, and MiniGPT-4 on a newly curated, culturally-rich Visual Question Answering (VQA) dataset specifically focused on traditional Korean attire, Hanbok. Our experimental results demonstrate that these models not only exhibit low accuracy but also reveal systematic error patterns indicative of a deeper lack of cultural understanding. Beyond diagnosing this deficiency, we propose a methodological refinement through the adoption of 'thick' evaluation frameworks that move beyond superficial accuracy metrics, explicitly assessing nuanced cultural understanding and alignment. Furthermore, we propose Multimodal Retrieval-Augmented Generation (MRAG) as an enhanced architectural paradigm to ground models explicitly in verifiable, culturally contextualized, and community-curated knowledge, addressing fundamental shortcomings of existing methods. This work provides empirical evidence of cultural limitations inherent in current VLMs and charts a research agenda toward building more equitable and culturally respectful AI for global digital heritage.

1. Introduction

Generative Artificial Intelligence is rapidly transforming the landscape of cultural heritage preservation, offering unprecedented tools for digital restoration, immersive educational experiences, and the democratization of access to global cultural treasures [20]. Vision-Language Models (VLMs), in particular, promise to interpret and articulate the rich narratives embedded in visual artifacts, from recreating lost historical sites to making archival materials accessible to a worldwide audience [3]. This technological promise, however, carries a significant risk: the potential for AI to become a medium of cultural homogenization rather than a guardian of diversity.

The core of this problem lies in the data that fuels these powerful models. State-of-the-art VLMs are predominantly trained on massive, web-scraped datasets that are heavily skewed towards Western, English-speaking, and high-income contexts [4, 16, 26]. This inherent data imbalance creates significant "cultural blind spots," leading to models that misinterpret, misrepresent, or are simply ignorant of artifacts from underrepresented cultures [24, 28, 30]. This is not merely a technical flaw; it is a situation that raises important ethical considerations and highlights the risk of cultural imbalance in algorithmic interpretation where dominant cultural norms become the default for AI-driven interpretation [13].

In this paper, we investigated this critical issue through the lens of Hanbok, the traditional attire of Korea. We found that leading VLMs such as LLaVA-1.5 [17], ViP-LLaVA [6], Shikra [9], and MiniGPT-4 [33] exhibit significant performance degradation when tasked with understanding the nuanced details of Hanbok. We used the Visual Question Answering (VQA) task as a diagnostic tool, as it moves beyond simple object recognition to probe for deeper contextual and compositional understanding [5]. Our experimental results show that these models not only fail in terms of accuracy but also exhibit specific, culturally revealing error patterns, such as conflating Hanbok with other East Asian garments (e.g., the Japanese kimono or Chinese hanfu) and generating factually incorrect details (hallucinations) about its components and cultural significance.

Our objective was to empirically highlight these failures and, more importantly, to chart a course for future research toward more culturally aware and equitable AI. The primary contributions of this paper are:

1. We identified and framed the problem of "cultural blind

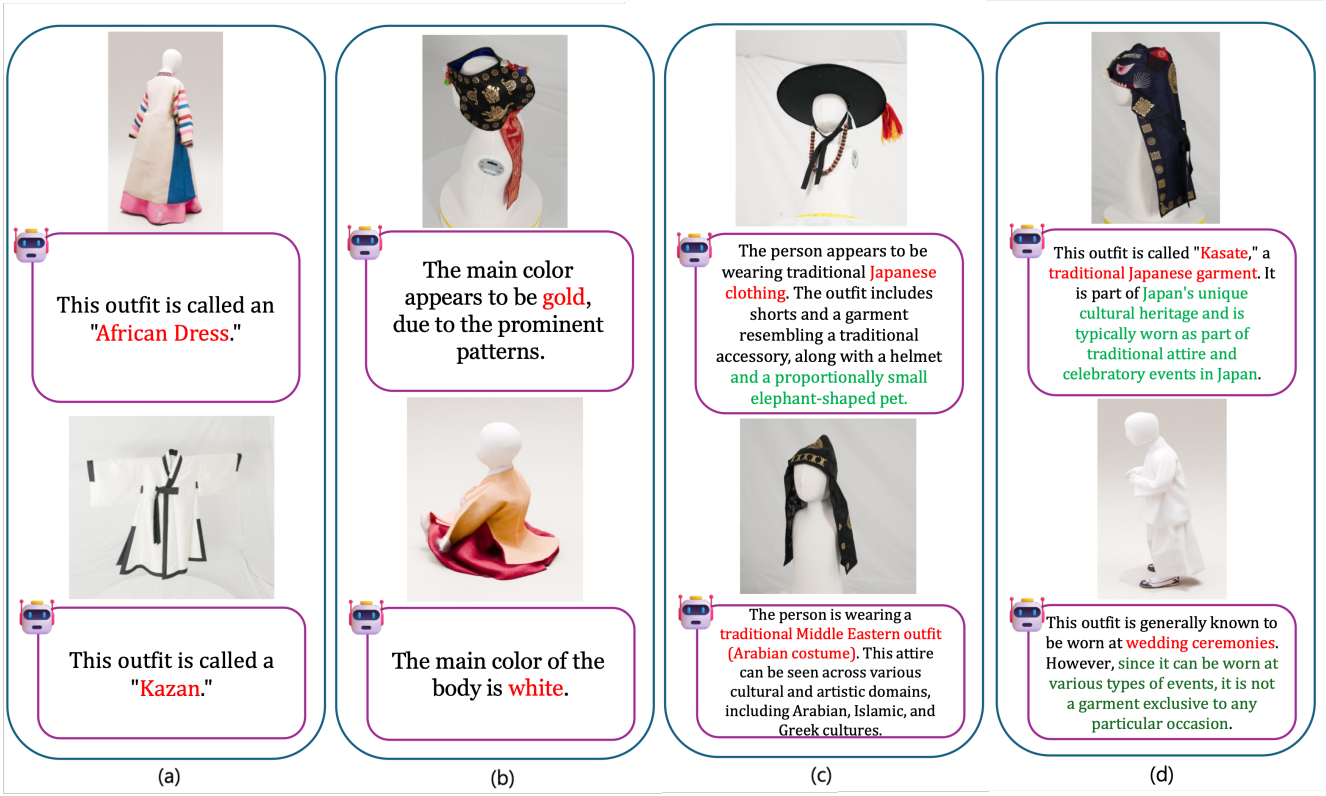

Figure 1. Failure (red) and hallucination (green) examples from VLM responses to Hanbok images. The input prompts are as follows: (a) level 1: "What is the name of this outfit?" (b) level 2: "What is the main color of the body?" (c) level 3: "Which country's traditional clothing is this person wearing?" (d) level 4: "What kind of events is this outfit typically worn for?"

spots" in contemporary VLMs, using the specific, high-impact case study of Korean Hanbok.

2. We implemented a comprehensive experimental framework, centered around a novel Hanbok-VQA dataset, to systematically evaluate the performance and failure modes of prominent open source VLMs.
3. We demonstrated a shift in evaluation, moving beyond simple accuracy to include qualitative analysis of culturally specific errors, such as cultural conflation, contextual blindness, and component hallucination.
4. We propose a forward-looking research agenda aimed at mitigating these biases, calling for the development of "thick" evaluation metrics co-designed with cultural communities [29] and the adoption of architectural solutions like Multimodal Retrieval-Augmented Generation (MRAG) to ground models in verifiable, curated cultural knowledge [22].

By demonstrating the limitations of current models and proposing concrete pathways forward, this paper aims to catalyze a critical conversation within the computer vision community. We argue that building culturally competent AI is not a niche concern but a fundamental requirement for the responsible and ethical development of technologies that will shape our collective digital future.

2. Related Work

2.1. Cultural Bias in Vision-Language Models

The problem of bias in large-scale AI models is well-documented [7, 12, 21, 23], and VLMs are no exception. Research showing that VLMs inherit and often amplify societal biases from their training data [4, 16, 19, 25]. The primary source of this bias is the composition of large-scale, web-scraped datasets like LAION and CC3M, which are heavily skewed towards Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies [27]. The common filtering practice favoring English-language image-text pairs further marginalizes non-Western and lower socioeconomic cultures [28]. Consequently, significant performance disparities have been observed, with models consistently achieving higher accuracy on Western cultural contexts compared to African or Asian contexts [18]. This "Western Gaze" [4] influences not only object recognition but also contextual understanding. For instance, VLMs trained on Western-centric datasets emphasize central objects while neglecting the background context crucial in East Asian de-

scriptive norms [13]. Empirical studies confirm these biases, revealing accuracy drops of up to 58% when evaluating models on culturally diverse images, highlighting a systematic cultural and ethnic bias [14].

2.2. Vision-Language Architectures and Instruction Tuning

Recent advancements in VLMs utilize visual instruction tuning, exemplified by architectures such as LLaVA and MiniGPT-4. LLaVA connects pre-trained vision encoders (e.g., CLIP ViT) with large language models (LLMs) like Vicuna, employing a two-stage training involving feature alignment and end-to-end instruction fine-tuning. MiniGPT-4 extends this by adding a second fine-tuning stage using high-quality datasets to improve natural language outputs. ViP-LLaVA facilitates interaction with visual prompts without complex region encodings, while Shikra incorporates referential dialogue, enabling spatially precise interactions via natural language. Our work evaluated these diverse architectures to examine their susceptibility to cultural bias stemming from varied vision-language integration approaches.

2.3. From Fine-tuning to Retrieval-Augmented Generation

Fine-tuning, a common method to adapt VLMs to specialized domains, faces limitations such as poor scalability across diverse cultures and "catastrophic forgetting [15]," where models lose general capabilities while overfitting specific tasks. Retrieval-Augmented Generation (RAG) offers a more scalable and robust alternative [11], enhancing language models by retrieving information from external knowledge bases, thereby reducing factual inaccuracies [22]. Multimodal RAG (MRAG) extends this approach to include images in the retrieval process, improving cultural accuracy and context-awareness without frequent retraining. MRAG provides greater transparency and updatability, making it ethically and technically superior for sensitive domains like cultural heritage. Tools such as LlamaIndex and Haystack facilitate the practical implementation of MRAG pipelines, validating the feasibility of this direction.

3. Probing Cultural Blind Spots: An Empirical Study on Hanbok VQA

To empirically validate the suspected cultural limitations of current Vision-Language Models, we conducted a systematic study on the task of Hanbok VQA. This section first establishes a framework for understanding the specific failure modes anticipated in this domain, then details our experimental protocol using a newly designed four-level question hierarchy, and finally presents a quantitative and qualitative analysis of our findings.

3.1. A Framework for Failure Analysis

We define three primary categories of errors that characterize a VLM's failure to comprehend culturally specific visual information:

- **Terminology Misidentification:** The inability to use precise, culturally specific terminology, often defaulting to generic Western analogues (e.g., identifying a 'goreum' as a 'ribbon').
- **Cultural Conflation:** A critical error where a model confuses one culture's artifacts with those of a geographically or culturally adjacent one (e.g., misidentifying Hanbok as a Japanese kimono).
- **Contextual Misunderstanding:** The failure to infer the social, historical, or situational context embedded in an attire, even if its basic visual attributes are recognized.

3.2. Experimental Protocol

To empirically test for the failures defined in our framework, we designed a rigorous experimental protocol covering dataset selection, question formulation, and evaluation methodology.

Dataset: The visual foundation of our study is the Traditional Korean Costume Image Dataset provided by Korea's AI-Hub [2]. This large-scale, high-quality dataset contains over 130,000 images of Hanbok, meticulously categorized by historical period, gender, occasion, and item type. From this rich collection, we curated a set of 100,000 Hanbok-VQA Probe Set. This set is a balanced subset of several hundred images ensuring diversity across styles (e.g., royal, commoner, ceremonial) and clarity of the main subject.

Question and Answer Formulation: For each image in our probe set, we manually crafted questions corresponding to our four-level cognitive hierarchy shown in Fig. 1: Level 1 (Identification), Level 2 (Attribute), Level 3 (Confusion), and Level 4 (Context). To ensure a fair and robust evaluation of the models' open-ended answers, we adopted a methodology inspired by the evaluation of non-binary questions in domain-specific datasets like Fashion-VQA [31].

For each question, we created a structured annotation containing a list of multiple acceptable ground-truth answers to account for linguistic and semantic variations. This allows for flexible matching beyond a single string comparison. An example of our annotation for a Level 3 question is shown in Tab. 1:

This structure ensures that if a model generates "Korea," "South Korea," or "Korean dress," it is recognized as a correct answer.

Evaluation Metric: The primary metric for our quantitative analysis is Top-1 Accuracy. A model's generated response for a given question is considered correct if the generated text contains an exact match to any of the strings

Field	Content
`question`	Which country's traditional clothing is this person wearing?
`gt_answers`	{"Korea", "South Korea", "Chosun", "Joseon", "Korean traditional clothing", "Korean traditional costume", "Korean dress", "Korean Hanbok"}

Table 1. Example of a culturally-aware VQA entry from the Hanbok-VQA dataset.

listed in the corresponding `gt_answers` array. This flexible matching approach allows us to fairly assess the semantic correctness of the models' free-form text outputs, moving beyond rigid string matching and better capturing the models' actual understanding. The accuracy is then calculated as the percentage of correctly answered questions across the entire probe set and within each of the four levels.

3.3. Results and Analysis

Our experimental results, presented in Tab. 2, confirm our central hypothesis: all tested models exhibit significant weaknesses in understanding Hanbok, and their performance systematically degrades as the required level of cultural reasoning deepens. The overall accuracy for all models languishes below 50%, highlighting a fundamental incompetence in this domain.

A nuanced analysis of the performance across levels provides deeper insights. Interestingly, all models performed slightly better on Level 2 (Attribute) questions than on Level 1 (Identification). This suggests that while the models possess competent foundational vision capabilities—they can correctly perceive objective attributes like color and texture they lack the specific cultural and lexical knowledge to name what they are seeing. For example, a model could correctly identify that a jeogori is 'red' (a Level 2 task) but fail to name it a 'jeogori', defaulting to 'a shirt' (a Level 1 failure).

Performance drops sharply at Level 3 (Confusion), validating our hypothesis of Cultural Conflation. With accuracies hovering in the mid-30s, the models frequently failed to distinguish Hanbok from the attire of neighboring countries, revealing a biased and poorly differentiated internal representation of East Asian cultures.

As predicted, the most profound failure occurs at Level 4 (Context), where accuracy plummets into the 20s. This highlights the models' near-complete inability to connect visual evidence to abstract cultural meaning. For instance, when presented with an image of sangbok (mourning wear), one of the top-performing models described it as "an elegant outfit suitable for a formal celebration," a response that is not only incorrect but culturally inappropriate.

These quantitative and qualitative results provide strong empirical support indicating that current VLMs exhibit notable cultural blind spots. Rather than reflecting only gaps in encyclopedic information, the patterns observed point to limitations in culturally grounded reasoning. These findings suggest a promising direction for future research, for which we propose a detailed roadmap in the following section.

4. A Roadmap for Culturally Aware VQA

The empirical evidence presented in our study (see Sec. 3) underscores the urgent need for a new approach to developing Vision-Language Models. The consistent failures in terminology, context, and factual grounding are not isolated errors but symptoms of a deeper, systemic issue rooted in data and architectural limitations. To address this challenge, we propose a comprehensive, three stage roadmap designed to guide future research toward building VLMs that are not just accurate, but genuinely culturally aware. This roadmap progresses from foundational data and model adaptation to advanced knowledge grounded reasoning and finally to a new paradigm for evaluation.

4.1. Stage 1: A Foundational Layer with Domain Specific Data and Adaptation

The most immediate bottleneck is the lack of high quality, culturally specific training data [13]. To overcome this, we first propose a systematic protocol for dataset creation and a practical method for model adaptation.

4.1.1. The Hanbok-VQA Dataset: A Hierarchical Protocol

We propose the creation of a new, large-scale Hanbok-VQA dataset. Drawing inspiration from methodologies used in domain specific VQA dataset creation such as FashionVQA [31], our protocol emphasizes a hierarchical question structure to ensure comprehensive coverage of knowledge, from basic identification to deep contextual reasoning. The questions should be categorized into four levels of increasing complexity:

- **Level 1 (Component Identification):** Questions about the names and colors of specific parts(e.g., "What is the color of the 'goreum' (고름, ribbon tie) on this 'jeogoril' (저고리, jacket)?").
- **Level 2 (Attribute Recognition):** Questions regarding materials, patterns, and production techniques (e.g., "Is this garment made of silk or ramie fabric?").
- **Level 3 (Contextual Understanding):** Questions about the social context, occasion, or status associated with the attire (e.g., "Is this outfit worn for a wedding ceremony or a funeral?").

Table 2. Quantitative comparison of VQA model performance on the Hanbok-VQA dataset (*hypothetical results*). Accuracy is reported overall and across four question levels: Level 1 (Identification), Level 2 (Attribute), Level 3 (Confusion), and Level 4 (Context).

Model	Overall Acc. (%)	Level 1	Level 2	Level 3	Level 4
ViP-LLaVA	41.9	48.0	52.5	38.1	28.9
LLaVA-1.5	40.9	47.3	51.8	37.2	27.4
MiniGPT-4	40.0	46.5	50.9	36.4	26.1
Shikra	39.0	45.2	50.1	35.5	25.3

- **Level 4 (Stylistic Reasoning):** Questions requiring historical and stylistic inference (e.g., "Does the silhouette of this attire reflect the style of the late Joseon dynasty?").

All data should be structured in a JSON format compatible with existing training pipelines, such as the conversations format used by LLaVA.

4.1.2. Parameter Efficient Domain Adaptation

With the dataset in place, the next step is to adapt pre-trained VLMs. Given the prohibitive cost of training from scratch, we advocate for the use of Parameter-Efficient Fine-Tuning (PEFT). Specifically, QLoRA (Quantized Low-Rank Adaptation) [10] offers a memory efficient solution to fine-tune large models on custom datasets with moderate hardware resources (e.g., a single A100 GPU). Frameworks like LLaMA-Factory [32] provide a unified and streamlined environment for applying QLoRA to over 100 different models, making this a highly practical and accessible starting point for domain adaptation.

4.2. Stage 2: Knowledge Intensive Reasoning with Multimodal RAG

Fine-tuning, a common method to adapt VLMs to specialized domains, faces limitations such as poor scalability across diverse cultures and "catastrophic forgetting," where models lose general capabilities while overfitting specific tasks [22]. Retrieval Augmented Generation (RAG) [8] offers a more scalable and robust alternative, enhancing language models by retrieving information from external knowledge bases, thereby reducing factual inaccuracies. Multimodal RAG (MRAG) [1] extends this approach to include images in the retrieval process, improving cultural accuracy and context awareness without frequent retraining. MRAG provides greater transparency and updatability, making it ethically and technically superior for sensitive domains like cultural heritage. Specifically, MRAG systems ensure answers are verifiable, significantly reduce hallucination by grounding responses in factual retrieval, and allow for scalable knowledge updates without costly retraining [11].

4.2.1. Proposed MRAG Architecture

We propose an architecture that is coordinated with a framework like LlamaIndex that supports practical implementations of MRAG pipelines. The pipeline is as follows:

- **Knowledge Base Curation:** A multimodal knowledge base is constructed from trusted sources (e.g., The National Folk Museum of Korea, Encyclopedia of Korean Culture). Each entry, or "knowledge unit," links descriptive text with high-resolution images of Hanbok and its components.
- **Multimodal Embedding:** The CLIP model is used to generate dense vector embeddings for both text and image data, projecting them into a shared semantic space.
- **Indexing:** These multimodal embeddings are stored and indexed in a specialized vector database such as ChromaDB or Milvus.
- **Retrieval Augmented Generation:** When a user poses a query with an image, the system embeds the query and retrieves the most relevant knowledge units (both text and images) from the vector DB. These retrieved artifacts are then prepended to the prompt of a VLM, providing rich, factual context to generate a grounded and accurate answer.

4.3. Stage 3: A New Paradigm for Evaluation

The ultimate success of a culturally aware VQA model cannot be measured by conventional metrics alone. Standard VQA accuracy, which treats all incorrect answers equally, is a "thin" evaluation method that fails to capture the rich, multi layered nature of cultural understanding. For instance, if a model identifies a 'goreum' (a traditional Korean coat string) as a 'ribbon', it might be considered functionally similar but is a profound cultural misinterpretation. An evaluation paradigm that cannot distinguish between this and a simple color misidentification is fundamentally inadequate for our purposes. Therefore, we propose a shift towards a "Thick Evaluation" framework [29] a concept that prioritizes qualitative depth and contextual nuance over simplistic, quantitative scores. As a concrete instantiation of this philosophy, we introduce the Cultural Accuracy Score (CAS), a composite metric designed to provide a holistic assessment of a model's cultural intelligence.

The CAS is composed of two primary dimensions: Quantitative evaluation of a model's explicit knowledge, and qualitative evaluation of its deeper, inferential understanding.

4.3.1. The Quantitative Dimension: Measuring Factual Knowledge

This dimension assesses the model's ability to correctly identify objective, verifiable facts from the image. It serves as a baseline for the model's perceptual and knowledge-retrieval capabilities.

Terminology F1-Score: This metric evaluates the model's command of the domain's specific vocabulary. A pre-defined list of essential Hanbok terms (e.g., gat, dongjeong, norigae) is used as a ground truth. We employ the F1-score, the harmonic mean of precision and recall, rather than simple accuracy. This is crucial because it penalizes both errors of commission (using a wrong term) and errors of omission (failing to use the correct term when appropriate), providing a more robust measure of terminological competence.

Attribute Accuracy: This metric functions closer to traditional VQA evaluation, measuring the model's accuracy on questions about explicit, objective visual attributes such as color, pattern, or material (e.g., "Is this skirt blue?"). This allows us to isolate the model's fundamental visual perception abilities from its deeper cultural reasoning.

4.3.2. The Qualitative Dimension: Assessing Inferential Understanding

This dimension is the core of the CAS, requiring human expert judgment to assess the model's nuanced understanding—an area where automated metrics fall short.

Contextual Appropriateness: This metric assesses whether the model's answer aligns with the socio-cultural, historical, and situational context of the attire shown. For example, an answer describing sangbok (mourning wear) as a "costume for a festive party" would receive the lowest possible score (e.g., 1 out of 5), even if its description of the material or color is factually correct. This evaluation must be conducted by a panel of experts in Korean history and costume, who rate each response on a pre-defined Likert scale.

Hallucination Rate: This is a critical metric for model trustworthiness. It measures the frequency with which a model fabricates information not present in the image. An expert annotator assigns a binary flag (1 for hallucination, 0 for none) to each response. An answer describing a non-existent phoenix embroidery on a plain garment is a critical failure. This metric is designed to heavily penalize models that are "confidently wrong," as such outputs are highly misleading.

4.3.3. Synthesizing the Composite Score

The final Cultural Accuracy Score is calculated as a weighted sum of these four components. The formula could be conceptualized as Eq. (1):

$$\begin{aligned} \text{CAS} = (\mathbf{w}_{\text{term}} \cdot \text{Terminology}_{\text{F1}}) + (\mathbf{w}_{\text{attr}} \cdot \text{Attribute}_{\text{acc}}) \\ + (\mathbf{w}_{\text{cont}} \cdot \text{Context}_{\text{score}}) - (\mathbf{w}_{\text{hall}} \cdot \text{Hallucination}_{\text{penalty}}) \end{aligned} \tag{1}$$

The weights ($\mathbf{w}$) must be carefully determined, reflecting the relative severity of each error type. For instance, a hallucination might be weighted most heavily as it represents a critical failure of factuality, while contextual understanding would also receive a high weight as it is central to cultural intelligence. While more labor intensive than traditional benchmarks, the CAS offers a high-resolution evaluation signal that can guide future improvements. It forces the research community to optimize not just for accuracy, but for nuance, context, and respect the true hallmarks of genuine understanding.

5. Conclusion

In this paper, we confronted the critical issue of cultural bias in modern Vision-Language Models through a focused case study on the Korean traditional attire, Hanbok. We have systematically demonstrated that current state-of-the-art models, despite their impressive capabilities on general benchmarks, consistently fail on nuanced, culturally specific queries, exhibiting predictable errors in terminology, contextual understanding, and factual grounding. In response to these validated shortcomings, we proposed a comprehensive, three-stage roadmap as a path forward for the research community. This roadmap advocates for the creation of rich, domain specific datasets, the adoption of more transparent and knowledge grounded architectures like Multimodal RAG, and a necessary shift towards a "Thick Evaluation" paradigm, exemplified by our proposed Cultural Accuracy Score (CAS). Ultimately, this work goes beyond identifying the limitations of current models; it aims to encourage further reflection and action. We argue that advancing visual intelligence requires a parallel commitment to cultural understanding. We encourage the computer vision community to look beyond optimizing for general benchmarks and to actively engage in the important work of creating AI systems that are more equitable, respectful, and truly representative of the diverse world they are meant to serve.

References

[1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025. 5

[2] AIHub. Traditional korean clothing image dataset (hanbok). https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71733, 2020. Accessed: 2025-07-09. 3

[3] Lafta Raheem Ali and Rihab Qassim Abdul-Kadim. The role of artificial intelligence in digitizing cultural heritage: A review. 1

[4] Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. See it from my perspective: How language affects cultural bias in image understanding. In *The Thirteenth International Conference on Learning Representations*. 1, 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[6] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vipllava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 1

[7] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*, 2023. 2

[8] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17754–17762, 2024. 5

[9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115, 2023. 5

[11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023. 3, 5

[12] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*, 2022. 2

[13] Ram Mohan Rao Kadiyala, Siddhant Gupta, Jebish Purbey, Srishti Yadav, Alejandro Salamanca, and Desmond Elliott. Uncovering cultural representation disparities in vision-language models. *arXiv preprint arXiv:2505.14729*, 2025. 1, 3, 4

[14] 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 bias of multimodal large language models in cultural mixture contexts. *arXiv preprint arXiv:2503.16826*, 2025. 3

[15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[16] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*, 2021. 1, 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[18] 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

[19] Zhaoming Liu. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244, 2025. 2

[20] Britney Johnson Mary. Generative ai for cultural heritage preservation. 1

[21] Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2023. 2

[22] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025. 2, 3, 5

[23] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023. 2

[24] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024. 1

[25] Joan Nwatu, Oana Ignat, and Rada Mihalcea. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. *arXiv preprint arXiv:2311.05746*, 2023. 2

[26] Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*, 2024. 1

[27] Uwe Peters and Mary Carman. Cultural bias in explainable ai research: A systematic analysis. *Journal of Artificial Intelligence Research*, 79:971–1000, 2024. 2

[28] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496, 2024. 1, 2

[29] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. The case for" thick evaluations" of cultural representation in ai. *arXiv preprint arXiv:2503.19075*, 2025. 2, 5

[30] 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. 1

[31] Min Wang, Ata Mahjoubfar, and Anupama Joshi. Fashion-vqa: A domain-specific visual question answering system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3514–3519, 2023. 3, 4

[32] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. 5

[33] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1