BLEnD-Vis: Benchmarking Multimodal Cultural Understanding in Vision Language Models

Bryan Chen Zhengyu Tan[1,2*] and **Weihua Zheng**[1,2*]
Zhengyuan Liu[2] **Nancy F. Chen**[2] **Hwaran Lee**[3] **Kenny Tsu Wei Choo**[1] **Roy Ka-Wei Lee**[1]
[1]Singapore University of Technology and Design (SUTD)
[2]Institute for Infocomm Research (I2R), A*STAR, Singapore
[3]Sogang University

Abstract

As vision-language models (VLMs) are deployed globally, their ability to understand culturally situated knowledge becomes essential. Yet, existing evaluations largely assess static recall or isolated visual grounding, leaving unanswered whether VLMs possess robust and transferable cultural understanding. We introduce **BLEnD-Vis**, a multimodal, multicultural benchmark designed to evaluate the *robustness* of everyday cultural knowledge in VLMs across linguistic rephrasings and visual modalities. Building on the BLEnD dataset, **BLEnD-Vis** constructs 313 culturally grounded question templates spanning 16 regions and generates three aligned multiple-choice formats: (i) a text-only baseline querying from Region → Entity, (ii) an inverted text-only variant (Entity → Region), and (iii) a VQA-style version of (ii) with generated images. The resulting benchmark comprises 4,916 images and over 21,000 multiple-choice questions (MCQ) instances, validated through human annotation. **BLEnD-Vis** reveals significant fragility in current VLM cultural knowledge; models exhibit performance drops under linguistic rephrasing. While visual cues often aid performance, low cross-modal consistency highlights the challenges of robustly integrating textual and visual understanding, particularly in lower-resource regions. **BLEnD-Vis** thus provides a crucial testbed for systematically analysing cultural robustness and multimodal grounding, exposing limitations and guiding the development of more culturally competent VLMs. Code is available at https://github.com/Social-AI-Studio/BLEnD-Vis.

1 Introduction

As large language models (LLMs) and vision-language models (VLMs) become increasingly embedded in global applications, their capacity to comprehend and respond to diverse cultural contexts is gaining critical importance (Pawar et al., 2024; Adilazuarda et al., 2024; Li et al., 2025b). While these models exhibit impressive general capabilities, they often falter in understanding everyday cultural practices—such as local foods, leisure activities, and family customs—particularly for communities that are underrepresented in mainstream training corpora (Myung et al., 2024; AlKhamissi et al., 2024). This poses real-world risks, as cultural insensitivity can undermine user trust, marginalise minority populations, and perpetuate global inequities in AI deployment (Qiu et al., 2025; Kannen et al., 2024).

Existing benchmarks that aim to evaluate cultural knowledge typically assess recall via direct textual prompts (Li et al., 2024a; Wang et al., 2024b) or focus on narrow multimodal contexts (Urailertprasert et al., 2024; Nayak et al., 2024; Satar et al., 2025; Winata et al., 2025). Yet, two essential questions remain underexplored: (1) How robust are these models to linguistic rephrasings of culturally grounded queries? (2) Can they consistently ground cultural knowledge in visual representations? These questions are important for distinguishing deep conceptual understanding from superficial or brittle associations that may degrade under linguistic or visual variation (Zhang et al., 2025; Lee et al., 2024).

To bridge this gap, we propose **BLEnD-Vis**, a multimodal, multicultural benchmark designed to evaluate the robustness and groundedness of everyday cultural knowledge in VLMs. **BLEnD-Vis** builds upon the BLEnD dataset (Myung et al., 2024) by selecting a curated set of tangible, culturally situated concepts across 16 diverse regions and constructing three parallel evaluation formats: *Original MCQ*, *Rephrased MCQ*, and *VQA-Style MCQ*. These formats enable controlled comparisons that isolate the effects of linguistic rephrasing and modality shifts on model performance. While this study focuses on English to leverage the strength of state-of-the-art models and ensure

[*] Equal contribution.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4647–4669

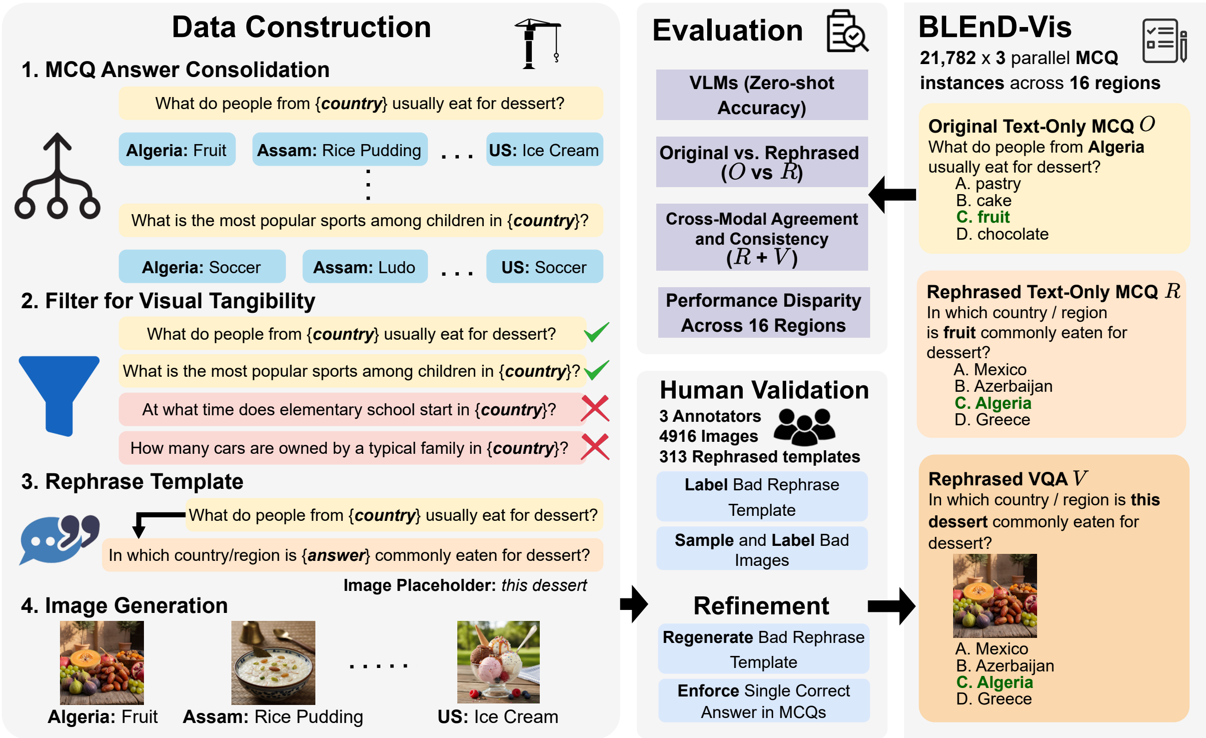


Figure 1: Overview of the **BLEnD-Vis** benchmark construction and evaluation framework. The process involves: (1) **Data Construction** via tangibility filtering, question rephrasing, and image generation based on BLEnD; (2) **Human Validation** of generated assets; (3) Creation of the final **BLEnD-Vis Dataset** comprising three parallel MCQ formats (Original Text, Rephrased Text, VQA) across 16 regions; and (4) **Evaluation** assessing VLM zero-shot accuracy, robustness to rephrasing, cross-modal consistency, and regional performance variations.

comparability across formats, **BLEnD-Vis** sets the stage for future multilingual expansion.

Our evaluation of twelve VLMs on **BLEnD-Vis** reveals that: (i) model performance does not strongly correlate with parameter count; (ii) linguistic rephrasing often degrades performance, indicating brittle knowledge representations; (iii) visual cues can significantly aid understanding, with VQA performance often surpassing text-only rephrased queries; and (iv) robust cross-modal consistency (correctness on both R and V formats for the same fact) is particularly challenging, with a mean joint correctness of only 42.19%. Furthermore, preliminary cross-modal fine-tuning experiments indicate that training on textual data generally improves VQA performance (mean +16.15%), while the transfer from VQA training to textual performance is less apparent (mean +3.98%). Figure 1 shows the overview of the BLEnD-Vis benchmark construction and evaluation framework. We summarise our contributions as follows:

- We present **BLEnD-Vis**, a benchmark to evaluate the robustness of everyday cultural knowledge in VLMs across linguistic and visual modalities.
- We develop a systematic pipeline involving tangibility filtering, question rephrasing, image generation, and human validation.
- We release a dataset of 313 validated templates, 4916 culturally grounded images, and 21,782 MCQs in three aligned formats.
- We provide a comparative evaluation of twelve VLMs, revealing gaps in cultural robustness and cross-modal generalisation.

BLEnD-Vis provides a rigorous framework to assess whether state-of-the-art VLMs encode not only cultural facts, but robust and transferable cultural understanding in both language and vision.

2 Related Works

2.1 Cultural Knowledge Benchmarks

Recent benchmarks assess cultural knowledge in LLMs through textual recall or alignment. BLEnD (Myung et al., 2024) captures everyday knowledge across regions and languages; CultureLLM (Li et al., 2024a), CulturePark (Li et al., 2024b), and CDEval (Wang et al., 2024b) measure alignment with cultural dimensions. SeaEval (Wang et al., 2024a) and MAPS (Liu et al., 2024a) extend this to multilingual or proverb-based reasoning. Palm (Alwajih et al., 2025) provides a culturally inclusive dataset but focuses on Arabic LLMs and textual modalities. RAVENEA (Li et al., 2025a) introduces a multimodal benchmark but prioritises Retrieval-Augmented Generation (RAG). While these benchmarks focus on cultural knowledge, they do not evaluate robustness to rephrasing or modality shifts. In contrast, **`BLEnD-Vis`** probes the robustness of cultural understanding through controlled textual rephrasing and image grounding, offering deeper diagnostic insights.

2.2 Cultural Bias and Alignment

Cultural bias in LLMs and VLMs is an ongoing concern, with studies showing that models often reflect Western-centric norms (Tao et al., 2024). Alignment methods have sought to mitigate this, usually using human-annotated survey responses as ground truth (AlKhamissi et al., 2024), or exposing covert harms through social scenario simulation (Dammu et al., 2024; Tan and Lee, 2025). Other work explores the causal assumptions embedded in culturally-influenced data (Bauer et al., 2023) and the fragility of model outputs when subjected to adversarial prompts (Zhang et al., 2025). Although these approaches surface latent biases, **`BLEnD-Vis`** complements them by testing representational stability across phrasings and visual formats, exposing brittle associations and performance disparities across diverse cultures.

2.3 Multimodal Cultural Evaluation

Multimodal benchmarks like SEA-VQA (Urailertprasert et al., 2024) and CulturalVQA (Nayak et al., 2024) test VLMs on culturally situated images, often exposing regional performance gaps. ALM-Bench (Vayani et al., 2025) broadens this across 100 languages. Others evaluate cultural competence in generative models (Kannen et al., 2024; D'Incà et al., 2024) or few-shot adaptation (Nikandrou et al., 2025; Kim et al., 2025). **`BLEnD-Vis`** is distinct in aligning textual and visual formats for the same cultural fact, enabling fine-grained comparisons and disentangling the effects of linguistic versus perceptual variation.

3 `BLEnD-Vis` Dataset Construction

The **`BLEnD-Vis`** benchmark is constructed via a multi-stage pipeline that transforms the textual, everyday cultural knowledge from BLEnD (Myung et al., 2024) into parallel evaluation sets suitable for probing textual and multimodal robustness.

3.1 Base Dataset and Scope

We utilise the 500 English Short Answer Question (SAQ) templates derived from the original BLEnD study as our starting point. The ground truth answers for these templates are sourced from the MCQ split of the BLEnD dataset[1] (Myung et al., 2024). Our current curation focuses on English to enable controlled comparisons across modalities and leverage state-of-the-art generation models.

3.2 Dataset Extension Pipeline

The pipeline proceeds through the following stages:

Step 1: MCQ Answer Consolidation. To associate each base SAQ template with its verified answer set across cultures, we processed the BLEnD MCQ dataset ($\sim$306k instances). For each MCQ instance, we extracted the correct answer text and its corresponding region ('*country*' field). This answer text was subsequently added to the known answers set for the specific question template '*ID*' and region, effectively consolidating all unique correct answers from the MCQ data for each base template across the 16 regions. Invalid answers (e.g., "*idk*" or "*i don't know*") were excluded.

Step 2: Tangibility Filtering. To ensure that questions are suitable for transformation into a VQA format, we automatically assessed the SAQ templates. Using GPT-4o (2024b) (prompt in Figure 8, Appendix F), we classified templates based on whether the question entailed answers that are **concrete, visually representable entities**. This filtering step selected 313[2] tangible question templates for further processing. For instance, in the example illustrated in Figure 1, the answer "*Rice Pudding*" is a visually representable entity.

Step 3: Question Rephrasing & Placeholder Generation. To create the textual condition for testing robustness to linguistic variation, we inverted the standard query format (Region $\rightarrow$ Entity) to

[1] `https://huggingface.co/datasets/nayeon212/BLEnD`, 'multiple-choice-questions' split.

[2] Initially, 320 templates were selected. However, during MCQ generation in Step 6, 7 templates lacked sufficient semantically distinct options to form valid 4-option MCQs, resulting in 313 usable templates.

(Entity $\rightarrow$ Region). For each of the 313 tangible question templates, GPT-4o was prompted (Prompt in Figure 9, Appendix F) to generate a canonical rephrased question template. Concurrently, the model generated a generic **image placeholder** (e.g., '*this food*') designed to replace the entity name in the VQA format, thereby compelling models to rely primarily on the visual modality for that task.

Step 4: Image Generation To enable multi-modal evaluations, we generated 4,916 culturally-contextualised images, one for each unique answer-region pair from the 313 tangible templates. We used Gemini 2.5 Flash Image (Fortin et al., 2025) for image generation, with prompts conditioned on the original question, specific answer, and region (prompt in Figure 10, Appendix F). To validate the use of synthetic images, we conducted a comparative study across three conditions (Old Synthetic, New Synthetic, and Human-Curated). As detailed in Appendix D.3, model performance on Gemini 2.5 Flash images showed a negligible difference (-1.7%) from human-curated real-world images, justifying their use as a scalable, high-fidelity proxy.

Step 5: Human Validation. A multi-stage validation process ensured the quality of all generated assets. First, three human annotators assessed the quality and semantic fidelity of 313 rephrased question templates, leading to the manual correction of 39 SAQ templates flagged by majority vote (see Appendix D for validation results, and Table 10 for examples of such corrections). Second, to validate the quality of the new Gemini 2.5 Flash image set, we designed a rigorous, sampled quality assurance protocol. A stratified random sample of 500 images ($\sim$10%) is evaluated by three annotators based on conceptual plausibility and recognisability (with instructions to use search engines to verify unfamiliar cultural concepts). This practical approach provides a statistical measure of dataset quality while remaining scalable. Full guidelines and validation details are provided in Appendix D.

Step 6: Parallel MCQ Generation. To create the final dataset, we generated three parallel MCQ sets for each core fact (validated tangible template + answer/region pair + corresponding image). For each fact, we generated up to 5 unique MCQ instances by sampling semantically distinct distractors (answers from different regions for the same template, with a simple substring check used to filter overly similar options). Uniqueness was enforced by tracking the set of answer options (1 correct, 3 distractors) for each fact and discarding duplicate sets of options. This yielded three parallel formats that tested the same knowledge point:

Original MCQ O: A text-only format querying cultural knowledge in the standard Region $\rightarrow$ Entity form.

Rephrased MCQ R: A linguistically inverted Entity $\rightarrow$ Region format testing sensitivity to phrasing variation.

VQA-Style MCQ V: A visual-grounded format pairing images with rephrased questions (Image + Placeholder $\rightarrow$ Region).

Examples of these parallel MCQ formats are provided in Appendix H (Table 18). To ensure the validity of the benchmark, a post-hoc uniqueness verification was performed to filter out any parallel set of MCQ instances where a distractor region could also be a valid answer for the queried entity in the R and V formats. This structured generation ensures diverse and non-repetitive test cases for robust cross-modal evaluation.

Category	MCQ Count	Percentage
Breakdown by Topic		
Education	1765	8.10 %
Family	2312	10.61 %
Food	6681	30.67 %
Holidays/Celebration/Leisure	4294	19.71 %
Sport	4650	21.35 %
Work life	2080	9.55 %
Breakdown by Country/Region		
Algeria (DZ)	1174	5.39 %
Assam (AS)	1761	8.08 %
Azerbaijan (AZ)	1180	5.42 %
China (CN)	1497	6.87 %
Ethiopia (ET)	1450	6.66 %
Greece (GR)	1449	6.65 %
Indonesia (ID)	1451	6.66 %
Iran (IR)	1331	6.11 %
Mexico (MX)	1522	6.99 %
North Korea (KP)	1287	5.91 %
Northern Nigeria (NG)[3]	998	4.58 %
South Korea (KR)	1532	7.03 %
Spain (ES)	1398	6.42 %
UK (GB)	1260	5.78 %
US	1296	5.95 %
West Java (JB)	1196	5.49 %
Total Instances	**21,782**	**100.00 %**

Table 1: **BLEnD-Vis**: MCQ Instance Count and Percentage Breakdown.

3.3 Resulting Dataset and Statistics

The final **BLEnD-Vis** benchmark comprises 313 tangible question templates drawn from BLEnD, each paired with a rephrased version and a corre-

[3]The lower instance count for Northern Nigeria is inherited from the upstream BLEnD data, which had higher rates of annotator 'I don't know' responses and refusals for this region.

sponding image placeholder to enable controlled evaluation across three modalities. In total, we generated 4,916 culturally grounded images, each corresponding to a unique answer-region pair. The question templates and culturally grounded images were used to construct 21,782 MCQ instances for each of the three parallel formats; each with the same topic and region distribution: *Original*, *Rephrased*, and *VQA-style*.

The three parallel formats of MCQs enable direct comparison of model performance on the same underlying cultural knowledge presented through different textual phrasings and modalities. For evaluations of cross-modal knowledge via unimodal training, the dataset is further split into training and test sets based on question templates to prevent data leakage (details in Appendix B, Table 7).

4 Results & Analysis

We evaluated 12 VLMs on the **BLEnD-Vis** benchmark to assess their robustness in representing everyday cultural knowledge. All evaluations were conducted in a zero-shot setting using the full dataset of 21,782 MCQ instances across three aligned formats. Models were prompted using a standardised evaluation template (Appendix F, Figure 11). Table 2 reports each model's accuracy on the three individual formats and includes two cross-modal consistency metrics. In particular, we sort models by their performance on the 'R-V Correct %' metric, which captures the percentage of cultural facts where the model correctly answered both the rephrased MCQ and the corresponding VQA-style MCQ, highlighting its ability to generalise consistently across modalities.

4.1 Overall Model Performance

Table 2 presents model performance across the three MCQ formats, revealing several notable trends in cultural knowledge robustness and multimodal reasoning.

Model size does not consistently predict performance. While performance generally scales within a model family (e.g., Qwen2.5-VL-32B outperforms the 7B variant on key consistency metrics like 'R-V Correct %'), performance across different model families does not strictly correlate with parameter count. For instance, smaller models like Kimi-VL-2.8B (48.51% 'R-V Correct %') and Llama-3.2-Vision-11B (48.01%) outperform the larger LLaVA-1.6-13B (41.03%). This suggests that factors beyond sheer parameter scale, such as the diversity of pre-training data, architectural choices for multimodal integration, and specific fine-tuning strategies, play a significant role in encoding robust cultural knowledge.

Rephrasing questions slightly reduces performance. Across models, average accuracy declines from 53.97% on the Original MCQ format to 52.03% on the Rephrased MCQ format. This reduced performance may stem from the prevalence of the (Region → Entity) format prevalent across many cultural benchmarks (Myung et al., 2024; Chiu et al., 2025; Romero et al., 2024). This implies that standard benchmark formats may overestimate a model's capability of cultural understanding, as the Entity → Region format may disrupt these learned patterns.

Visual input provides important cultural cues. Models perform better on the VQA format (69.82%) than on both the Rephrased text-only format (52.03%) and the Original text-only format (53.97%). For instance, Kimi-VL-2.8B improves from 52.22% (Rephrased) to 83.21%(VQA), and Qwen2.5-VL-7B from 49.06% to 84.91%. This is likely because images can convey additional culture-specific cues for cultural-knowledge retrieval while textual prompts might be informationally sparse. However, this implies that evaluating cultural capabilities through VQA alone can be misleading, as high VQA scores may not reflect true multimodal reasoning and may mask underlying weaknesses under unimodal text-only settings.

Cross-modal consistency remains a challenge. Despite improved accuracy in the VQA format, models struggle to produce consistent, correct answers across modalities. The average 'R-V Agree %', the proportion of matched predictions across the Rephrased and VQA formats, is moderately high at 62.53%, but the stricter 'R-V Correct %', accuracy on both formats simultaneously, drops to 42.19%. Even the top-performing model (GPT-4o) achieves only 60.83% on 'R-V Correct %'. This gap highlights that many correct answers in one modality are not replicated in the other, reflecting cross-modal fragility. This motivates future work on modal consistency, highlighting the need for benchmarks, evaluations, and training processes that optimise for robust cross-modal grounding over unimodal accuracy.

Model	**Original MCQ**	**Rephrased MCQ**	**VQA MCQ**	**R-V Agree (%)**	**R-V Correct (%)**
GPT-4o (2024b)	**69.56**	**63.36**	**92.01**	<u>66.29</u>	**60.83**
Qwen2.5-VL-32B (2025)	<u>61.90</u>	<u>57.32</u>	<u>86.03</u>	63.83	<u>53.59</u>
Kimi-VL-2.8B (2025)	57.13	52.22	83.21	61.59	48.51
Llama-3.2-Vision-11B (2024a)	58.45	54.57	81.24	60.21	48.01
Qwen2.5-VL-7B (2025)	58.05	49.06	84.91	58.19	46.08
Molmo-7B-D (2024)	53.99	50.57	72.39	62.82	42.89
InternVL3-8B (2025)	57.18	54.68	64.07	65.79	42.27
LLaVA-1.6-13B (2024b)	44.89	50.85	67.65	63.56	41.03
LLaVA-1.6-7B (2024b)	41.40	46.05	63.22	65.53	37.40
PaliGemma2-10B (2024)	54.26	52.64	54.35	**67.37**	37.18
DeepSeek-VL2-small-2.8B (2024)	50.76	49.43	45.95	54.40	24.89
NVILA-2B (2025)	40.10	43.58	42.78	60.77	23.57
Mean (Overall)	**53.97**	**52.03**	**69.82**	**62.53**	**42.19**

Table 2: Zero-Shot Accuracies (%) of VLMs on `BLEnD-Vis` (Full Dataset). 'R-V Agree %' indicates the percentage of instances where the model's prediction for the **Rephrased** MCQ matched its prediction for the **VQA** MCQ. 'R-V Correct %' indicates the percentage of instances where the model answered *both* the **Rephrased** and **VQA** MCQs correctly. Models ordered by 'R-V Correct %'. Best **bolded**, second best <u>underlined</u>.

4.2 Performance Variation by Region

Table 3 reports mean accuracy across all models for each cultural region, aggregated over the three MCQ formats. Figure 2 visualises VQA-style performance per region and model, highlighting distinct regional strengths and weaknesses. Additional breakdowns for the Original and Rephrased text-only formats are included in Appendix A.2.

Region	**Original**	**Rephrased**	**VQA**	**Mean**
US	64.92	**77.42**	80.87	**74.40**
UK (GB)	**65.97**	69.95	81.15	72.36
China (CN)	61.87	67.76	**82.20**	70.61
South Korea (KR)	58.56	66.99	78.18	67.91
Mexico (MX)	59.12	55.00	77.11	63.74
Spain (ES)	58.17	55.20	72.85	62.07
Indonesia (ID)	52.84	59.01	73.48	61.78
Greece (GR)	54.39	48.42	74.57	59.13
Iran (IR)	52.13	46.27	70.27	56.22
Azerbaijan (AZ)	54.13	43.47	65.56	54.39
Northern Nigeria (NG)	44.41	48.33	67.10	53.28
West Java (JB)	44.16	44.29	59.27	49.24
Ethiopia (ET)	43.79	37.45	64.67	48.64
Assam (AS)	46.63	37.97	57.07	47.22
Algeria (DZ)	50.74	36.75	53.34	46.94
North Korea (KP)	49.46	35.48	55.46	46.80
Mean (Regions)	**53.83**	**51.86**	**69.57**	**58.42**

Table 3: Model Accuracies (%) by Country/Region on `BLEnD-Vis` (Full Dataset), ordered by mean task performance.

Model performance varies significantly by region, often reflecting resource disparities. Regions with greater representation in publicly available training data tend to yield higher model performance. For example, the US, UK, and China achieve average accuracies of **74.40%**, **72.36%**, and **70.61%**, respectively. In contrast, less digitally represented regions such as North Korea, Algeria, and Assam score markedly lower, with mean accuracies of **46.80%**, **46.94%**, and **47.22%**. This trend aligns with prior findings from BLEnD (Myung et al., 2024) and underscores persistent gaps in cultural representation within pretraining corpora (Tao et al., 2024).

The (Entity → Region) query format exacerbates performance gaps between high and low-resource regions. For the **Original MCQ**, the performance range between the top-performing region (UK: **65.97%**) and a bottom-performing region (e.g., Ethiopia: **43.79%**) is approximately 22%. This gap significantly widens for the **Rephrased MCQ** (US: **77.42%** vs. North Korea: **35.48%**) and the VQA format (China: **82.20%** vs. Algeria: **53.34%**). This suggests that when cued with an entity (textually or visually) and asked for its associated region, models may default to high-resource regions if their knowledge of the entity's specific origin in a low-resource region is less robust. This suggests that the (Entity → Region) format, particularly in VQA, is a more discriminative test of potential regional biases and deep cultural grounding. For instance, as illustrated in Figure 2, models like Qwen2.5-VL-7B exhibit strong VQA performance for China (95.26%), while their scores for a region like Algeria are much lower (67.38%). Conversely, models like PaliGemma2-10B and Molmo-7B-D, while generally strong for US VQA (64.04% and 85.03% respectively), show comparatively lower performance for China (61.72% and 83.58% respectively).

4.3 Cross-Modal Knowledge Transfer

To explore whether cultural knowledge learned in one modality can transfer to another, we conducted preliminary cross-modal fine-tuning experiments on a subset of models. In the first set-

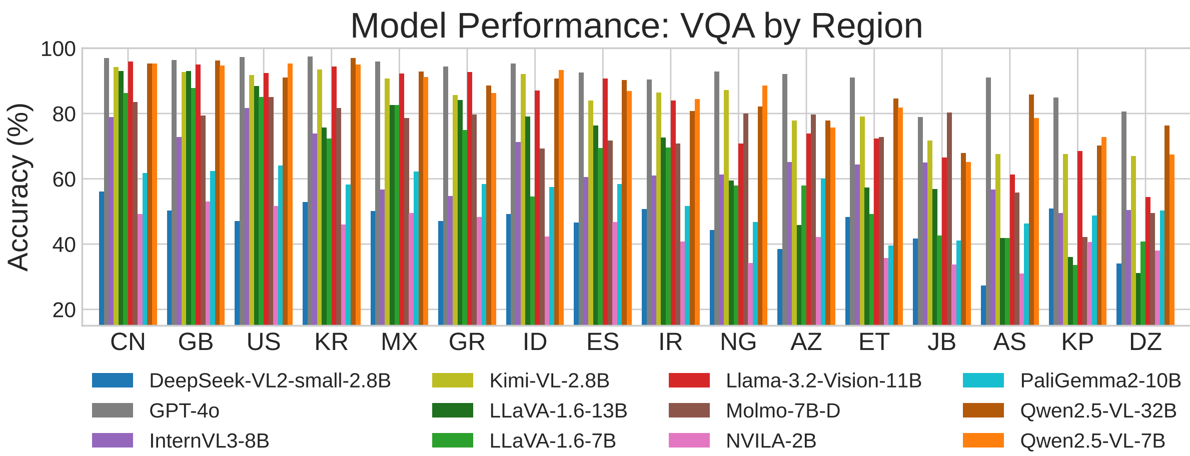


Figure 2: Accuracies (%) of each evaluated VLM for the VQA-Style MCQ format in **BLEnD-Vis** (Full Dataset) across 16 different cultural regions (see Appendix A.1, Table 6 for region code definitions), highlighting regional variations in model performance. (Original and Rephrased text-only formats are in Appendix A.2.)

ting (Text-trained → VQA-test), models were fine-tuned on the training split of the Rephrased Text-Only MCQs and evaluated on the test split of the VQA-Style MCQs. In the reverse setting (VQA-trained → Text-test), models were fine-tuned on VQA-Style MCQs and evaluated on Rephrased Text-Only MCQs. In both cases, we used the same 80/20 template-based train/test split to prevent leakage of cultural facts between training and evaluation (Appendix B). Full fine-tuning hyperparameters are provided in Appendix C.

Target Task: Rephrased Text (Test Set)		
Model	**Performance (%)**	
	Baseline	**VQA-Trained**
LLaVA-1.6-7B	44.22	52.82 (+8.60 %)
Qwen2.5-VL-7B	47.83	51.66 (+3.83 %)
PaliGemma2-10B	51.39	53.72 (+2.33 %)
Llama-3.2-Vision-11B	51.73	52.91 (+1.18 %)
Mean (Models)	**48.79**	**52.78 (+3.98 %)**

Target Task: VQA (Test Set)		
Model	**Performance (%)**	
	Baseline	**Text-Trained**
LLaVA-1.6-7B	63.63	78.35 (+14.72 %)
Qwen2.5-VL-7B	85.25	91.06 (+5.81 %)
PaliGemma2-10B	52.67	92.60 (+39.93 %)
Llama-3.2-Vision-11B	80.64	84.78 (+4.14 %)
Mean (Models)	**70.55**	**86.70 (+16.15 %)**

Table 4: Cross-Modal Transfer Performance (% Accuracy on Test Set). 'Baseline' refers to zero-shot performance on the target task. 'VQA-Trained' and 'Text-Trained' refer to performance on the target task after fine-tuning on VQA or Rephrased Text training data, respectively. Percentage improvement over baseline shown in parentheses.

Table 4 summarises these transfer learning results, detailing performance on the target task after fine-tuning on the source task, compared to their zero-shot baselines. The results indicate varying degrees of knowledge transfer across modalities.

Notably, fine-tuning on the Rephrased Text-Only MCQs consistently leads to improved performance on the VQA-Style MCQs (Text-trained → VQA-test). All four evaluated models show gains in this direction: LLaVA-1.6-7B (+14.72%), Qwen2.5-VL-7B (+5.81%), PaliGemma2-10B (+39.93%), and Llama-3.2-Vision-11B (+4.14%). On average, text-based training boosted VQA performance by **+16.15%**. This suggests that strengthening the model's textual understanding of the (Entity → Region) relationship and associated cultural facts can positively transfer to its ability to perform the same task when presented with visual cues, potentially by solidifying semantic representations that the visual modality can then leverage more effectively.

Conversely, the transfer from VQA training to Rephrased Text-Only performance (VQA-trained → Text-test) is more modest: with LLaVA-1.6-7B (+8.60%), Qwen2.5-VL-7B (+3.83%), PaliGemma2-10B (+2.33%), and Llama-3.2-Vision-11B (+1.18%) all showing lower improvements. The average improvement across these models is modest at **+3.98%**. This might suggest that while VQA training exposes models to visual-textual pairings of cultural concepts, it may not consistently enhance (and could potentially interfere with) purely textual reasoning pathways. It is plausible that VQA training could lead to an over-

Topic	Original	Rephrased	VQA	Mean
Work life	**67.34**	**66.04**	71.15	**68.18**
Holidays/Celeb.	56.15	55.18	**71.38**	60.90
Sport	55.32	52.04	71.28	59.55
Food	52.90	50.27	67.14	56.77
Education	45.90	44.57	70.67	53.71
Family	44.44	44.31	69.88	52.88
Mean	**53.68**	**52.07**	**70.25**	**58.66**

Table 5: Mean Model Performance (%) by Topic on **BLEnD-Vis** (Full Dataset).

reliance on visual features or that the VQA task structure does not reinforce nuanced textual understanding as effectively as direct textual training.

These findings highlight the interplay between modalities in representing and reasoning about cultural knowledge. The positive transfer from text to VQA is promising, suggesting robust textual understanding as foundational. However, the less consistent transfer from VQA to text warrants further investigation into how multimodal training influences distinct reasoning pathways.

4.4 Performance Variation by Topic

The distribution of mean model performance by topic (Table 5) reveals substantial variation in task difficulty across cultural domains. Models performed best on '*Work life*' (mean accuracy: 68.18%) and '*Holidays/Celebration/Leisure*' (60.90%), while topics such as '*Family*' (52.88%) and '*Education*' (53.71%) posed greater challenges. These differences may reflect varying levels of specificity, visual distinctiveness, or semantic ambiguity associated with the entities in each topic.

Across all categories, performance was consistently higher on the VQA format compared to the Rephrased text-only format. For example, accuracy on 'Education' increased from **44.57%** (Rephrased) to **70.67%** (VQA), and on '*Family*' from **44.31%** to **69.88%**. These results suggest that visual input provides a valuable disambiguating signal, particularly for topics where textual rephrasings may be less canonical or culturally entangled. The consistent VQA boost underscores the utility of grounded visual context in supporting flexible retrieval of everyday cultural knowledge.

5 Discussion

Our findings highlight several challenges in current VLMs' representation of everyday cultural knowledge. First, the consistent drop in accuracy under linguistic rephrasing suggests that many models rely on superficial pattern matching rather than robust conceptual understanding. While visual input in the VQA format generally improves performance, low cross-modal consistency (particularly in joint correctness) reveals a lack of integration between textual and visual representations.

Regional disparities further underscore equity concerns: models perform significantly worse on queries involving lower-resource regions, a gap potentially exacerbated by limitations in VLM training data and the cultural fidelity of generated images. This motivates greater inclusivity in pretraining data and culturally-aware generative tools.

Notably, model scale does not reliably predict performance. Instead, our results suggest that the diversity and specificity of pretraining data, along with architectural design, may be more critical to cultural robustness. The observed asymmetry in cross-modal transfer, where text-based fine-tuning enhances VQA performance but not vice versa, reinforces the foundational role of linguistic grounding in multimodal understanding.

Together, these insights call for a shift in VLM development: beyond factual recall and scale, toward deeper, transferable, and culturally representative knowledge across modalities. Future training strategies could explicitly up-weight underrepresented cultural samples to counteract high-resource bias, and include structurally diverse templates (such as inverted Entity $\rightarrow$ Region formats) to improve linguistic robustness.

6 Conclusion & Future Work

We introduced **BLEnD-Vis**, a multimodal benchmark for evaluating the robustness and visual grounding of everyday cultural knowledge in vision-language models. Covering 16 culturally diverse regions, the benchmark includes 313 question templates, 4,916 generated images, and over 21,000 MCQ instances across three formats. Evaluations of 12 VLMs reveals key limitations: (i) performance degrades under linguistic rephrasing, cross-modal consistency remains low, and regional disparities persist for underrepresented cultures. (ii) Model scale does not reliably predict success, while fine-tuning results suggest that strong textual grounding supports more effective visual transfer. Future work includes expanding to multilingual settings, analysing failure patterns, improving culturally-aware training and generation methods, and extending evaluations to open-ended tasks.

Limitations

While BLEnD-Vis provides a novel framework for evaluating cultural robustness in VLMs, several limitations should be acknowledged. First, the benchmark is constructed entirely in English, limiting its applicability to multilingual and cross-lingual settings and inherently framing cultural concepts through an Anglophone lens. Future extensions should explore culturally grounded evaluation in other languages and code-mixed contexts. Second, the VQA component relies on synthetically generated images. Although human validation was performed, these images may still exhibit subtle inaccuracies or stereotypical cues, influenced by the biases of the image generation models. This could affect the fairness and fidelity of the VQA evaluation, particularly for underrepresented cultures. Furthermore, while our validation confirms synthetic images are a valid performance proxy, they may "flatten" cultural nuances compared to real artifacts (e.g., depicting generic rather than region-specific variants). Third, BLEnD-Vis focuses on tangible, everyday cultural knowledge. More abstract cultural dimensions, such as values, norms, or social rituals—are not represented, leaving a gap in evaluating deeper forms of cultural competence. Fourth, the use of MCQs limits evaluation to discriminative reasoning. While suitable for controlled comparisons, MCQs do not capture generative abilities such as explaining cultural facts, expressing empathy, or engaging in culturally appropriate open-ended dialogue. As such, the benchmark does not reflect models' full potential in real-world, interactive scenarios. Fifth, the current VQA format (Image + Placeholder $\rightarrow$ Region) tests only one type of visual grounding. Alternative visual query structures could expose different strengths or weaknesses in multimodal reasoning. Lastly, while human validation was conducted, annotator familiarity may have varied across the 16 regions. For less globally prominent cultures, subtle inaccuracies or overlooked errors may persist. These limitations point to important directions for future work, including multilingual expansion, generative evaluation, culturally adaptive image synthesis, and a broader model evaluation landscape.

Ethics Statement

The development and deployment of BLEnD-Vis were guided by a commitment to responsible research practices. We acknowledge several ethical considerations inherent in evaluating cultural knowledge in AI models.

Bias and Representation: The benchmark aims to *reveal* potential biases in VLMs concerning cultural knowledge, particularly disparities between high-resource and lower-resource regions, as evidenced in our results. However, the benchmark itself could inadvertently perpetuate biases present in the original BLEnD data or introduced during image generation (e.g., stereotypical depictions), despite human validation efforts. By making the benchmark public, we intend to facilitate research into identifying and mitigating such biases, promoting more equitable cultural representation in future models. We focused on everyday cultural knowledge to minimise the risk of evaluating sensitive or sacred topics inappropriately.

Data Provenance and Annotation: The textual data originates from the BLEnD dataset (Myung et al., 2024), which involved human participants providing cultural information. Consistent with the original BLEnD dataset's licensing, we release our dataset under a CC-BY-4.0 license. Images were generated using publicly available models and subsequently validated by human annotators following defined guidelines (Appendix D) to ensure relevance and appropriateness, filtering out problematic content flagged by a majority vote. Annotators are focused on validation rather than subjective judgment of cultural value. Furthermore, AI assistants provided support for coding tasks and enhancing the clarity of manuscript drafts; all such contributions were meticulously reviewed and edited by the authors to ensure the final work's accuracy and adherence to academic standards.

Intended Benefit: Our primary goal is to contribute positively to the AI community by providing a tool that encourages the development of VLMs with a more nuanced, robust, and equitable understanding of diverse global cultures. We believe that improving the cultural competence of AI systems is crucial for fostering user trust, reducing harm caused by cultural insensitivity, and promoting fairness in global AI applications. The dataset and associated code will be released publicly to facilitate transparency and further research in this critical area.

Acknowledgements

This research project is supported by the National Research Foundation, Singapore, under its National

Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004 and AISG-NMLP-2024-005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation and Ministry of Education, Singapore. We also sincerely thank Yuriel Ryan, Yujia Hu, Dylan Raharja and Zhuoran Wang for their contributions with annotation and evaluation.

References

Meta AI . 2024a. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

OpenAI . 2024b. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. 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. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL technical report. https://arxiv.org/abs/2502.13923v1.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanushree Mitra. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. 2024. OpenBias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235.

Alias Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. 2025. Introducing gemini 2.5 flash image, our state-of-the-art image model- google developers blog. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji B. Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Advances in Neural Information Processing Systems*, 37:13716–13747.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. 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. Association for Computational Linguistics.

Tony Lee, Haoqin Tu, Chi H. Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S. Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. 2024. VHELM: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. CulturePark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. 2025a. RAVENEA: A benchmark for multimodal retrieval-augmented visual culture understanding.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025b. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. 2025. NVILA: Efficient frontier visual language models.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating In-context adaptation of vision and language models to culture-specific concepts. 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 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark.

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

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. PaliGemma 2: A family of versatile VLMs for transfer.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. 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 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. 2025. Kimi-VL technical report.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

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. 2025. All languages matter: Evaluating LMMs on culturally diverse 100 languages.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, 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 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. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, Nanning Zheng, and Kaipeng Zhang. 2025. B-AVIBench: Toward evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *Trans. Info. For. Sec.*, 20:1434–1446.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models.

A Detailed Performance Breakdowns

This appendix provides further detailed breakdowns of model performance.

A.1 Region Code Mapping

Table 6 lists the 16 cultural regions included in the **BLEnD-Vis** benchmark and their corresponding two-letter codes used internally and in some data representations.

Code	Country / Region Name
AS	Assam
AZ	Azerbaijan
CN	China
DZ	Algeria
ES	Spain
ET	Ethiopia
GB	United Kingdom (UK)
GR	Greece
ID	Indonesia
IR	Iran
JB	West Java
KP	North Korea
KR	South Korea
MX	Mexico
NG	Northern Nigeria
US	United States (US)

Table 6: Mapping of Region Codes to Country/Region Names.

A.2 Performance by Region and Model (Text-Only Formats)

Figures 3 and 4 illustrate the performance of each model on the Original Text-Only MCQ and Rephrased Text-Only MCQ formats, respectively, broken down by cultural region.

B Dataset Split Details

The **BLEnD-Vis** dataset, comprising 21,782 MCQ instances derived from 313 unique question template IDs, was partitioned into training and test sets. The split was performed at the template ID level using an 80%-20% ratio (250 IDs for training, 63 IDs for testing) with stratification based on the topic

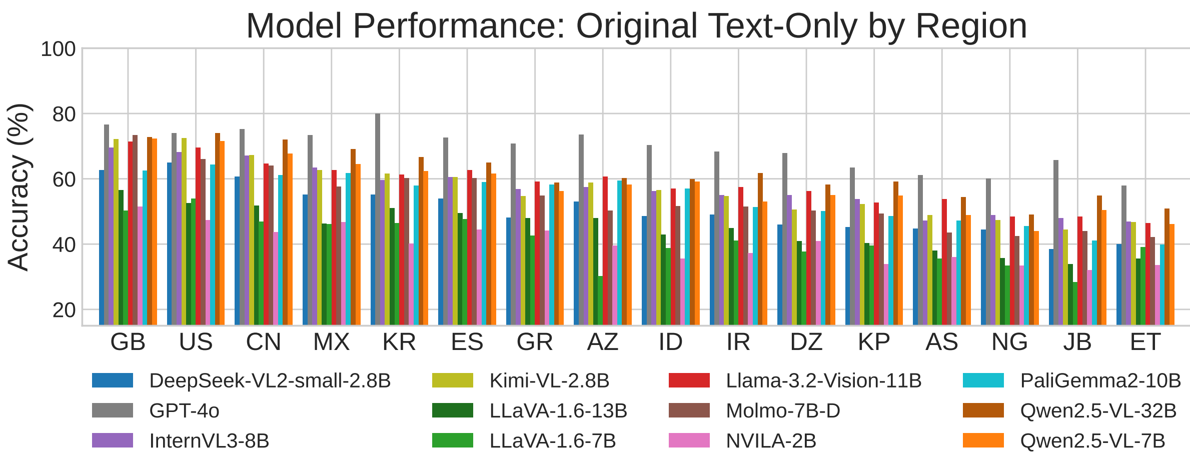


Figure 3: Original Text-Only MCQ Performance (%) by Region and Model on **BLEnD-Vis** (Full Dataset).

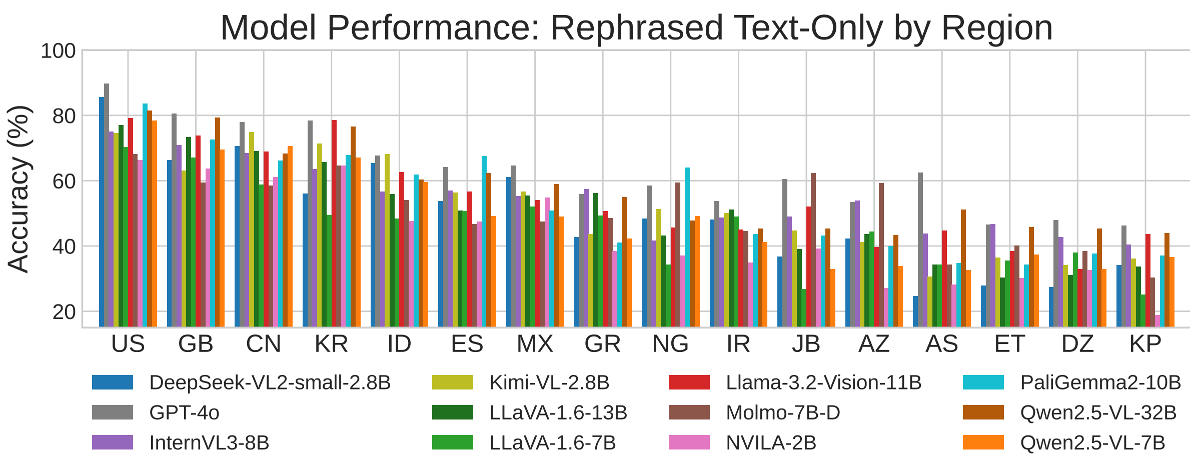


Figure 4: Rephrased Text-Only MCQ Performance (%) by Region and Model on **BLEnD-Vis** (Full Dataset).

category. This ensures that all MCQ instances originating from the same base template reside in the same split, preventing data leakage. Table 7 details the resulting distribution of MCQ instances across topics for the training and test sets.

C Fine-tuning Details for Cross-Modal Transfer Experiments

This section outlines the experimental setup for the cross-modal transfer learning experiments discussed in Section 4.3.

C.1 Dataset and Splitting

All fine-tuning and evaluation for the transfer experiments were conducted using the **BLEnD-Vis** dataset. We utilised the predefined train-test split detailed in Appendix B, which partitions the 313 unique question template IDs into an 80% training set (250 IDs; 17,320 MCQs) and a 20% test set (63 IDs; 4,462 MCQs). This split ensures that no underlying cultural facts or question templates seen during training are present in the test set, preventing data leakage.

Two primary training scenarios were investigated:

1. **Text-trained → VQA-test:** Models were fine-tuned on the Rephrased Text-Only MCQs from the training split and subsequently evaluated on the VQA-Style MCQs from the test split.

2. **VQA-trained → Text-test:** Models were fine-tuned on the VQA-Style MCQs from the training split and subsequently evaluated on the Rephrased Text-Only MCQs from the test split.

Zero-shot baseline performance was established by evaluating the pre-trained models directly on

Topic	Split	Count	Percentage
Education	Total	1765	8.1 %
	Train	1366	7.9 %
	Test	399	8.9 %
Family	Total	2312	10.6 %
	Train	1823	10.5 %
	Test	489	11.0 %
Food	Total	6681	30.7 %
	Train	5302	30.6 %
	Test	1379	30.9 %
Holidays/Celeb.	Total	4294	19.7 %
	Train	3512	20.3 %
	Test	782	17.5 %
Sport	Total	4650	21.3 %
	Train	3575	20.6 %
	Test	1075	24.1 %
Work life	Total	2080	9.5 %
	Train	1742	10.1 %
	Test	338	7.6 %
Overall	Total	**21782**	**100.0 %**
	Train	**17320**	**100.0 %**
	Test	**4462**	**100.0 %**

Table 7: Topic Distribution in Total, Train, and Test Splits of **BLEnD-Vis** MCQs.

the respective test splits without any **BLEnD-Vis** specific fine-tuning.

C.2 Fine-tuning Hyperparameters

The fine-tuning process for both LLaVA-1.6-7B and Qwen2.5-VL-7B utilised a consistent set of key hyperparameters, aiming for a standardised comparison. LoRA (Hu et al., 2021) was employed for efficient fine-tuning, targeting all linear layers. The models were trained for 3 epochs, and the checkpoint corresponding to the best validation loss (or training loss if a separate validation set was not carved out from the training split for hyperparameter tuning) was selected for final evaluation. Table 8 lists the pertinent hyperparameters.

C.3 Model Size and Computational Resources

All experiments were conducted utilising NVIDIA H100 GPUs (80GB). For inference, a single GPU with a batch size of 1 was employed. Vision-Question Answering (VQA) tasks required approximately 1.5 to 2.5 hours per model for the full dataset, and around 10 minutes for split dataset evaluations. Text-only task inference was faster, taking 30 to 50 minutes for the full dataset and approximately 10 minutes for split dataset runs per model. Model training was performed on a distributed setup of 8 NVIDIA H100 GPUs, with an average training duration of approximately 1 hour per model configuration.

Hyperparameter	Value
Fine-tuning Type	LoRA
LoRA Rank (r)	8
LoRA Target Modules	All linear layers
Learning Rate	1.0e-4
Number of Train Epochs	3.0
LR Scheduler Type	Cosine
Warmup Ratio	0.1
Batch Size (per device)	8
Gradient Accumulation Steps	8
Mixed Precision	bf16
Optimizer	AdamW
Weight Decay	0.01
Max Sequence Length	4000 (cutoff_len)

Table 8: Key Fine-tuning Hyperparameters.

D Human Annotation and Image Validation

To ensure the quality of all generated assets and to validate our use of synthetic images, we conducted a multi-stage human validation process. This involved (1) validating all rephrased question templates, (2) performing sampled quality assurance on the main image dataset, and (3) conducting a direct comparative study of our synthetic images against a human-curated baseline. All annotators are research assistants with at least an undergraduate degree.

D.1 Rephrased Question Validation

Three independent annotators evaluated the 320 automatically generated rephrased question templates for semantic fidelity and clarity against the original BLEnD SAQ templates. The goal was to verify if the rephrased question accurately inverted the original query while maintaining clarity.

Results: A majority vote (at least 2 of 3 annotators) flagged 39 templates (12.2%) as 'Bad', often due to semantic misalignment or grammatical issues. These were manually corrected by the research team. An additional 7 templates were later excluded due to insufficient distractor options, resulting in the 313 validated templates used for the final MCQ dataset.

D.2 Sampled Quality Assurance for Generated Images

All 4,916 images in **BLEnD-Vis** were generated using the Gemini 2.5 Flash model. To quantify the quality of this large dataset, we implemented a sampled quality assurance protocol.

Task Setup: A random, stratified sample of 500 images ($\sim$10%) was evaluated by three independent annotators. The annotators' goal was to determine if each generated image served as a plausible and recognisable visual representation of its intended cultural concept, based on the entity, region, and a descriptive placeholder.

Annotation Guidelines: Annotators were provided with a detailed protocol to determine if a generated image serves as a plausible and recognisable visual representation of a specific cultural concept. For each image, annotators were given three key pieces of context: the specific ‘entity‘ (e.g., "spicy potatoes"), the ‘region‘ (e.g., "Spain"), and the general category or ‘image_placeholder‘ (e.g., "this food").

The core instructions were as follows:

- **Mark as ’G’ (Good) if:** The image successfully conveys the intended concept. The guiding principle was whether someone familiar with the region's cultural context would likely understand that the image represents the intended ‘entity‘. Images were marked ’G’ even with minor flaws, such as:
  - Slightly unusual art styles (e.g., painterly or illustrative), as long as the subject was identifiable.
  - Minor visual artefacts or imperfections (e.g., odd textures, background glitches).
  - Representations that might seem generic in isolation but were appropriate within the given cultural context.
- **Mark as ’B’ (Bad) if:** The image fails to sufficiently represent or convey the core concept. Annotators were required to provide a brief reason for their judgment. The primary criteria for a ’B’ rating were:
  - *Clearly Wrong Subject:* The image depicts something fundamentally different from the ‘entity‘.
  - *Misleading or Ambiguous Representation:* The image is so generic, abstract, or poorly rendered that it fails to represent the ‘entity‘, even with the provided context.
  - *Unusably Distorted:* The image has such severe visual artefacts that the main subject is unrecognisable or nonsensical.

Annotators were also encouraged to use image search engines to verify concepts with which they were unfamiliar, ensuring a high degree of accuracy in their judgments.

Results: Based on a 2/3 majority vote, **27/500 (5.40%)** of the sampled images were flagged as ’B’ (Bad). This low error rate suggests a strong overall quality and recognisability for the Gemini 2.5 Flash image set, supporting its suitability for our main evaluations.

D.3 Validation Against Human Curation

A key concern for benchmarks using synthetic data is whether the data serves as a valid proxy for real-world scenarios. To address this directly, we conducted a controlled experiment comparing VLM performance on our new synthetic images against both an older generation of synthetic images and a newly created, human-curated set of real-world images.

Methodology:

1. We randomly sampled a set of 100 cultural facts from our benchmark, disjoint from the human quality assurance sample.
2. For these 100 facts, we created three parallel sets of images:
   - **Synthetic (2.0):** Images generated using an older model (Gemini 2.0 Flash Image).
   - **Synthetic (2.5):** The final images used in our benchmark, generated by Gemini 2.5 Flash Image.
   - **Human-Curated:** Real-world images manually sourced by annotators via web search engines to best represent the target concepts. To ensure strict compliance with intellectual property standards, these images were used exclusively for this internal comparative experiment and are not included in the public benchmark release. This allows us to validate performance against real-world data while ensuring the released dataset remains open-source and license-free.
3. We ran VQA evaluations for a subset of models across these three parallel image sets.

Results and Justification: Table 9 presents the comparative VQA performance. The results show that **model performance on the new Gemini 2.5 Flash images is nearly identical to performance on the human-curated images**, with a mean difference of -1.7%. In contrast, the older synthetic images (Gemini 2.0 Flash) resulted in a significant performance drop of over 21.3% compared to the human-curated set. This experiment provides strong evidence that the high-fidelity Gemini 2.5 Flash images serve as a valid proxy for real-world data in our evaluation context, ensuring our results are robust and representative of real-world visual understanding challenges.

Model	Human Curated	Synth (2.0) (vs. Human)	Synth (2.5) (vs. Human)
Qwen2.5-VL-32B	83.00	63.00 (-20.0)	81.00 (-2.0)
Llama-3.2-Vision-11B	78.00	55.00 (-23.0)	76.00 (-2.0)
LLaVA-1.6-7B	68.00	47.00 (-21.0)	67.00 (-1.0)
Mean (Overall)	**76.33**	**55.00 (-21.3)**	**74.67 (-1.7)**

Table 9: VQA performance comparison. Parentheses show the performance change (% absolute) of synthetic images relative to the human-curated baseline.

E Annotation Examples: Rephrased Questions and Images

This section provides illustrative examples of issues identified during the human annotation phase (Task 1: Rephrased Question Validation and Task 2: Image Validation), highlighting common failure modes in automated generation and the importance of the validation step.

E.1 Examples of Corrected Rephrased Questions

During validation, 39 rephrased question templates were flagged by a majority of annotators as 'BAD' due to semantic misalignment or lack of clarity. These were manually corrected. Table 10 presents examples of original templates, their initially generated (problematic) rephrasings, and the manually corrected versions used in the final dataset.

The initial rephrasings often failed to capture the specific nuance of the original question (e.g., focusing only on celebration rather than family association or religious nature) or created grammatically awkward structures. The manual corrections aimed to restore the original intent while adhering to the required (Entity $\rightarrow$ Region) format.

E.2 Examples of 'BAD' Images

Validation also identified images that failed to accurately or appropriately represent the target concept. Figures 5 through 7 illustrate examples flagged as 'BAD' by all three annotators, showcasing common failure modes such as ambiguous representation, severe generation artefacts, and a lack of regional specificity.

Figure 5: **ID:** Ca-sp-45, **Topic:** Family, **Region:** Iran, **Target Answer:** 'north'.
Reason Flagged 'BAD': The image of a family in a forest is too generic and lacks specific visual cues to represent a destination in the 'north' of Iran, failing to convey the intended concept. Failure mode: *Ambiguous/Unclear Representation*.

These examples highlight ongoing challenges in automated image generation, particularly in creating images that are not only free of artefacts but also visually specific and culturally/geographically accurate. The validation step was crucial for identifying such failures.

ID	Original Template	Initial (BAD) Rephrasing	Corrected (GOOD) Rephrasing
Ji-ko-25	What are the family-related holidays in {country}?	*Template:* During which occasion is {answer} celebrated? *Placeholder:* this celebration	*Template:* In which country/region is {answer} the holiday most associated with family? *Placeholder:* this holiday
Sa-en-22	What is the most famous university in {country} known for its sports team?	*Template:* At which university is {answer} known for its sports team? *Placeholder:* this university	*Template:* {answer} is the most famous university known for its sports team in which country/region? *Placeholder:* this university
New-as-01	What is the most famous religious holiday in {country}?	*Template:* During which occasion is {answer} celebrated? *Placeholder:* this religious holiday	*Template:* In which country/region is {answer} the most famous religious holiday? *Placeholder:* this holiday

Table 10: Examples of Manually Corrected Rephrased Question Templates.

Figure 6: **ID:** Th-en-03, **Topic:** Sport, **Region:** Ethiopia, **Target Answer:** ‘football league’.
Reason Flagged ’BAD’: The image contains severe generation artefacts, such as the player’s lower body being clipped by the ground, which makes it look unnatural and distorted. Failure mode: *Unusably Distorted.*

Figure 7: **ID:** Na-ko-45, **Topic:** Holidays/Celebration/Leisure, **Region:** Azerbaijan, **Target Answer:** ‘qabala’.
Reason Flagged ’BAD’: The generated landscape is too generic and does not accurately reflect the visual characteristics of the actual destination, Qabala in Azerbaijan. Failure mode: *Inaccurate Regional Representation.*

F Prompts Used in Dataset Curation and Evaluation

This section details the prompts provided to Large Language Models (LLMs) and image generation models during the automated stages of the `BLEnD-Vis` dataset construction pipeline.

F.1 Tangibility Classification Prompt (GPT-4o)

Purpose: Classify if a question template and its answers refer to tangible concepts suitable for image generation (Figure 8). Used in Step 2.

F.2 Question Rephrasing & Placeholder Generation Prompt (GPT-4o)

Purpose: Rephrase the original question template to invert the query (Entity → Region) and generate a generic placeholder text for the VQA-style format (Figure 9). Used in Step 3.

F.3 Image Generation Prompt (Gemini-Flash-2.5)

Purpose: Generate a culturally contextualised image representing a specific answer entity, using the original question for context (Figure 10). Used in

User Prompt:

```
Please determine if the following question and its
answers across different regions can be visually
represented in an image.

Question: {question_template}

Answers across regions:
{formatted_answers}

Task: Analyse whether this question and its
answers reference tangible concepts that can be
clearly depicted in an image. Also analyse whether
answers reference specific entities (people or place).

Classification criteria examples:
- Tangible (include): food, drinks, items, sports,
occupations, festivals, commodities, famous people,
infrastructure, religious symbols, instruments,
clothes, animals, common physical activities
- Intangible (exclude): dates/times, ages, musical
genres, languages, software, numbers, insurance,
abstract concepts, entrance exams, education subjects

Also label whether any answer references a
specific person or place.

For consistency:
1. A question about "What is the most popular X" can
be tangible if X itself can be visually depicted
2. Questions about specific quantities (e.g., "How
many hours...") are generally intangible
3. Questions about time periods, ages, or numeric data
are intangible
4. Consider the question AND the answers - all must be
visually representable

Please respond with:
- is_tangible: true/false
- reason: brief explanation for your decision
- specific_entity: true/false (whether any answer
references a specific person or place)

Format your response in JSON:
{
"is_tangible": boolean,
"reason": "string",
"specific_entity": boolean
}
```

Figure 8: Prompt used for Tangibility Filtering (Step 2).

Step 4.

F.4 VLM Evaluation Prompt Template

Purpose: General template used to query Vision-Language Models for all three MCQ formats (Original Text, Rephrased Text, VQA-Style) in `BLEnD-Vis`. For VQA-Style, an image is provided to the model preceding this textual prompt (Figure 11).

User Prompt:

```
Your task is to rephrase the original question template
so that the concept represented by the original
answers becomes the new subject. The goal is to create
a single, natural-sounding rephrased question template
(using an {answer} placeholder) and a single generic
image placeholder text. This rephrased template and
placeholder should work well regardless of which
specific answer (from the examples below or similar
ones) is eventually used. The rephrased question
should implicitly ask for the country/region where the
{answer} fits the description, and it must NOT contain
any country/region information itself.

Original question template:
'{question_template}'

This template can produce questions for different
countries, with answers such as: {sample_answers_str}.

Based on the original template and the type of
answers it elicits, please generate:
1. A rephrased question template. It MUST include the
exact placeholder {answer}. It should be phrased to
ask 'In which country/region...?' or similar, making
the location the expected answer.
2. A short, generic text (under 10 words) to be used
as a placeholder when an image of the answer is shown
alongside the rephrased question template. This text
should describe the *type* of thing the answer is
(e.g., 'this food', 'this item', 'this tradition',
'this landmark', 'this event').

Separate the rephrased template and the image
placeholder text with the special sequence:
[IMAGE_PLACEHOLDER_SEP]

Do not add explanations. Ensure the {answer}
placeholder is present in the rephrased template.

Example 1:
Original question template: 'What is a popular street
food snack in {country}?'
Sample answers: 'churros', 'takoyaki', 'pretzels'
Your response should be:
In which country/region is {answer} a popular street
food snack?
[IMAGE_PLACEHOLDER_SEP]
this street food

Example 2:
Original question template: 'What is the name of the
traditional clothing worn by women in {country}?'
Sample answers: 'kimono', 'sari', 'dirndl'
Your response should be:
In which country/region is {answer} the traditional
clothing worn by women?
[IMAGE_PLACEHOLDER_SEP]
this traditional clothing

Example 3:
Original question template: 'What is a common way to
greet someone in {country}?'
Sample answers: 'bowing', 'handshake', 'wai'
Your response should be:
In which country/region is {answer} a common way to
greet someone?
[IMAGE_PLACEHOLDER_SEP]
this greeting

Now, based on the provided original question
template and answer examples, generate the rephrased
template and image placeholder.

Rephrased question template and image placeholder:
```

Figure 9: Prompt used for Question Rephrasing & Placeholder Generation (Step 3).

```
User Prompt:
{question_template_instance}

Given the above question, generate a photorealistic
and natural-looking photo of the following:
photo of {original_answer} in {country}. 4k HDR
beautiful, natural, Macro, High detail, focus,
full-color.
```

Figure 10: Prompt used for Image Generation (Step 4).

```
User Prompt:
{formatted_question_text}
Without any explanation, choose only one from the
given alphabet choices(e.g., A, B, C). Provide as JSON
format: {{"answer_choice":""}}

A. {choice_A_text}
B. {choice_B_text}
C. {choice_C_text}
D. {choice_D_text}

Answer:
```

Figure 11: General prompt template used for VLM evaluation across all **BLEnD-Vis** MCQ formats. For VQA, an image precedes this text.

G Analysis of Cross-Modal Agreement Patterns

To provide deeper insight into model behavior, we analysed the patterns of agreement and disagreement between the Rephrased text-only (R) and VQA (V) formats across all models. We categorised each instance into one of five outcomes. The quantitative breakdowns by topic and region are presented, followed by qualitative examples illustrating each pattern.

G.1 Quantitative Analysis of Agreement Patterns

Tables 11 and 12 show the distribution of outcomes. The analysis reveals two key patterns:

Systematic Bias in Low-Resource Regions: Models often share the same incorrect answer across both text and vision for low-resource regions, suggesting entrenched biases. The 'Agree & Incorrect' rate for regions like **North Korea (8.13%)** and **Assam (6.87%)** is over **3 times higher** than for the **US (2.00%)**. This indicates that when models are uncertain, they converge on the same plausible (but wrong) high-resource answer in both modalities.

Topic-Specific Modality Strengths: In the topic breakdown, **Work life** has the highest 'Agree & Correct' rate (**51.71%**), showing strong cross-modal understanding. However, it also has the highest 'Disagree (R_Correct)' rate (**14.33%**). This suggests that concepts related to "Work life" (e.g., specific job roles or workplace norms) are well-represented textually but can be visually ambiguous or difficult to depict in a single image, causing the VQA modality to fail more often.

Topic	Agree-Corr	Agree-Incorr	Disagree (R✓)	Disagree (V✓)	Disagree-Incorr
Food	40.87	5.33	9.40	26.27	18.13
Sport	42.56	3.97	9.48	28.71	15.27
Hols./Celeb.	44.98	4.09	10.20	26.40	14.34
Family	35.98	4.74	8.33	33.90	17.06
Work life	51.71	2.96	14.33	19.44	11.55
Education	36.30	4.10	8.27	34.36	16.97

Table 11: Cross-Modal Outcome Patterns by Topic (%).

Region	Agree-Corr	Agree-Incorr	Disagree (R✓)	Disagree (V✓)	Disagree-Incorr
Assam	27.99	6.87	9.98	29.09	26.08
South Korea	56.07	2.54	10.92	22.11	8.37
Mexico	46.57	3.64	8.43	30.54	10.82
China	59.42	2.41	8.34	22.78	7.05
Indonesia	48.30	3.60	10.71	25.18	12.22
Ethiopia	29.28	4.82	8.18	35.40	22.33
Greece	40.84	4.49	7.57	33.73	13.37
Spain	45.79	3.99	9.42	27.06	13.75
Iran	35.93	4.37	10.35	34.34	15.01
US	66.34	2.00	11.09	14.53	6.04
North Korea	25.08	8.13	10.41	30.38	26.00
UK	60.63	2.57	9.31	20.52	6.96
West Java	33.96	6.37	10.33	25.31	24.03
Azerbaijan	33.44	4.34	10.03	32.12	20.07
Algeria	25.38	6.64	11.38	27.96	28.64
North. Nigeria	35.93	3.92	12.40	31.17	16.58

Table 12: Cross-Modal Outcome Patterns by Region (%).

G.2 Qualitative Examples of Agreement Patterns

We present five examples from the Qwen2.5-VL-32B model to illustrate each cross-modal outcome. Each example shows the model's response to both text-only (Rephrased) and image-based (VQA) versions of the same question.

MCQ Info & Choices	Model Responses (Qwen2.5-VL-32B)
MCQID: Th-en-15_18 **Topic:** Sport **Entity:** hockey **Choices:** A. Assam, **B. US (Correct)**, C. UK, D. Ethiopia	***Rephrased (Text-Only):*** In which country/region is hockey the most popular winter sport? **Answer:** B. US (Correct) ***VQA (Image):*** In which country/region is this winter sport the most popular winter sport?  **Answer:** B. US (Correct)

Table 13: Agree & Correct – The model correctly answers in both modalities, demonstrating robust, well-grounded knowledge.

MCQ Info & Choices	Model Responses (Qwen2.5-VL-32B)
MCQID: Jo-sp-02_65 **Topic:** Sport **Entity:** chess **Choices:** A. Mexico, B. Northern Nigeria, **C. Greece (Correct)**, D. North Korea	***Rephrased (Text-Only):*** In which country/region is chess the most popular sport played without a ball? **Answer:** D. North Korea (Incorrect) ***VQA (Image):*** In which country/region is this sport the most popular sport played without a ball?  **Answer:** C. Greece (Correct)

Table 14: Disagree (VQA Corrects Text) – The model fails textually but the visual cue helps it recover the correct answer, highlighting the value of grounding.

MCQ Info & Choices	Model Responses (Qwen2.5-VL-32B)
MCQID: New-en-59_13 **Topic:** Family **Entity:** russia **Choices:** A. UK, B. Algeria, **C. Azerbaijan (Correct)**, D. South Korea	***Rephrased (Text-Only):*** In which country/region is russia the most popular destination for families to emigrate to? **Answer:** C. Azerbaijan (Correct) ***VQA (Image):*** In which country/region is this destination the most popular destination for families to emigrate to?  **Answer:** A. UK (Incorrect)

Table 15: Disagree (Text Corrects VQA) – The model succeeds textually but fails with the visual input.

MCQ Info & Choices	Model Responses (Qwen2.5-VL-32B)
MCQID: Th-en-15_1 **Topic:** Sport **Entity:** skiing **Choices:** **A. UK (Correct)**, B. Assam, C. US, D. North Korea	***Rephrased (Text-Only):*** In which country/region is skiing the most popular winter sport? **Answer:** C. US (Incorrect) ***VQA (Image):*** In which country/region is this winter sport the most popular winter sport?  **Answer:** C. US (Incorrect)

Table 16: Agree & Incorrect – The model provides the same incorrect answer in both modalities, indicating a shared, systematic bias.

MCQ Info & Choices	Model Responses (Qwen2.5-VL-32B)
MCQID: Jod-ch-15_10 **Topic:** Food **Entity:** snail **Choices:** A. South Korea, B. US, **C. Assam (Correct)**, D. UK	***Rephrased (Text-Only):*** In which country/region is snail a popular type of seafood? **Answer:** D. UK (Incorrect) ***VQA (Image):*** In which country/region is this seafood a popular type of seafood?  **Answer:** A. South Korea (Incorrect)

Table 17: Disagree & Both Incorrect – The model is incorrect in both modalities and provides different answers, suggesting a lack of knowledge.

H Examples of Parallel MCQ Formats

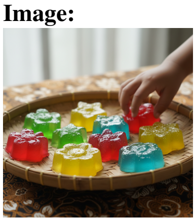
MCQ-ID	MCQ Format	Question & Options
Al-en-01_1	Original (Region → Entity)	**Q:** What is a common snack for preschool kids in West Java? **Options:** A. toast B. candy C. mashed potato rice D. jelly
	Rephrased (Entity → Region)	**Q:** For which country/region is jelly a common snack for preschool kids? **Options:** A. Greece B. North Korea C. Assam D. West Java
	VQA-Style (Image → Region)	**Image:**  **Q:** For which country/region is this snack a common snack for preschool kids? **Options:** A. Greece B. North Korea C. Assam D. West Java
Th-en-01_9	Original (Region → Entity)	**Q:** What is the most popular summer sport in Ethiopia? **Options:** A. volleyball B. running C. swimming D. badminton
	Rephrased (Entity → Region)	**Q:** In which country/region is running the most popular summer sport? **Options:** A. Spain B. Ethiopia C. Azerbaijan D. Indonesia
	VQA-Style (Image → Region)	**Image:**  **Q:** In which country/region is this sport the most popular summer sport? **Options:** A. Spain B. Ethiopia C. Azerbaijan D. Indonesia

Table 18: Examples of Parallel MCQ Formats in `BLEnD-Vis`.