BANGLAPROTHA: Evaluating Vision Language Models in Underrepresented Long-tail Cultural Contexts

Md Fahim[1,2]*, Md Sakib Ul Rahman[3]*, Akm Moshiur Rahman[2]*, Md Farhan Ishmam[4]*,
Md Tasmim Rahman[1], Fariha Tanjim Shifat[1], Fabiha Haider[1], Md Farhad Alam Bhuiyan[1]

[1]*Penta Global Limited* [2]*CCDS, Independent University, Bangladesh*
[3]*University of Maryland, Baltimore County* [4]*University of Utah*

*Equal Contribution {fahimcse381, souroveskb, pdcsedu}@gmail.com

Abstract

The advanced multimodal processing of current vision language models (VLMs) has prompted rigorous benchmarking across multicultural settings, revealing a clear inclination toward Western culture. While the bias likely stems from the predominance of Western-centric images in the VLM pretraining data, the resulting long-tail distribution problem is only exacerbated in underrepresented cultural settings, such as Bengali. Our work explores this problem through an aspect-based evaluation of several classes of VLMs on the rich Bengali culture. Our BanglaProtha dataset is a VQA dataset, containing images that encapsulate Bengali cultural elements, questions in native Bengali, and semantically similar multiple-choice answer options. Our experiments provide behavioral insights into VLMs across prompting & fine-tuning strategies, cultural aspects, model size, and augmentation methods. Our work serves as a diagnostic tool for addressing and mitigating inequalities in multicultural and multilingual settings, thereby bringing efforts to democratize AI systems. Our code and data are available at https://github.com/farhanishmam/BanglaProtha.

1. Introduction

The recent scaling of VLMs [10, 50] led to extensive improvement across several vision-language tasks, such as visual question answering (VQA) [7], visual grounding [37], and visual reasoning [28]. The visio-linguistic elements of the associated datasets vary widely, thereby challenging the exceptionally advanced systems [25]. One such variation arises in multicultural settings where the images contain cultural artifacts, *e.g.*, region, event, architecture, *inter alia*, that are typically not observed in benchmark datasets [45]. The associated questions also center on these cultural aspects and are often multi-lingual, where non-English scripts bring linguistic variations in the textual modality [53].

There have been several instances where VLMs exhibit substantial bias towards Western cultural concepts [45, 52]. The performance disparity can be attributed to the imbalance in pre-training data of the associated VLMs [47] and hence re-framing the multi-cultural evaluation problem to a long-tail distribution problem, *i.e.*, the niche cultural elements are present at the tail-end of the pre-training data distribution. As contemporary VLMs often require or rely on a training paradigm [24], the lack of resources makes it challenging to mitigate this skewed distribution problem.

Despite recent interest in multicultural and multilingual VLM evaluation [9, 53], the cultural depth of such benchmarks remains shallow. We exemplify this through the diverse Bengali[1] culture. To characterize the uniqueness of this culture, several aspects must be considered, *e.g.* food, events, landmarks, and art. However, existing benchmarks rarely explore beyond surface-level aspects of this culture [45]. Additionally, Bengali has a unique blend of Indic and Southeast Asian cultural elements, often requiring a multi-faceted understanding of cultural concepts [43].

With several culture-specific benchmarks in other non-Western cultures [46, 61], we found it crucial to construct a dataset and evaluate VLMs on culturally unique aspects of Bengali. Our contributions can be summarized as:

- We present BANGLAPROTHA, a VQA dataset with images encapsulating nine distinct Bengali cultural aspects, questions in native Bengali, and semantically similar multiple-choice answer options.
- We evaluate monolingual, multilingual, and large-scale VLMs on our dataset using five prompting and four fine-tuning strategies to assess their performance.

[1]In this work, *Bangla* and *Bengali* are used synonymously to denote the same language, cultural identity, and people, predominantly associated with Bangladesh and the West Bengal region of India.

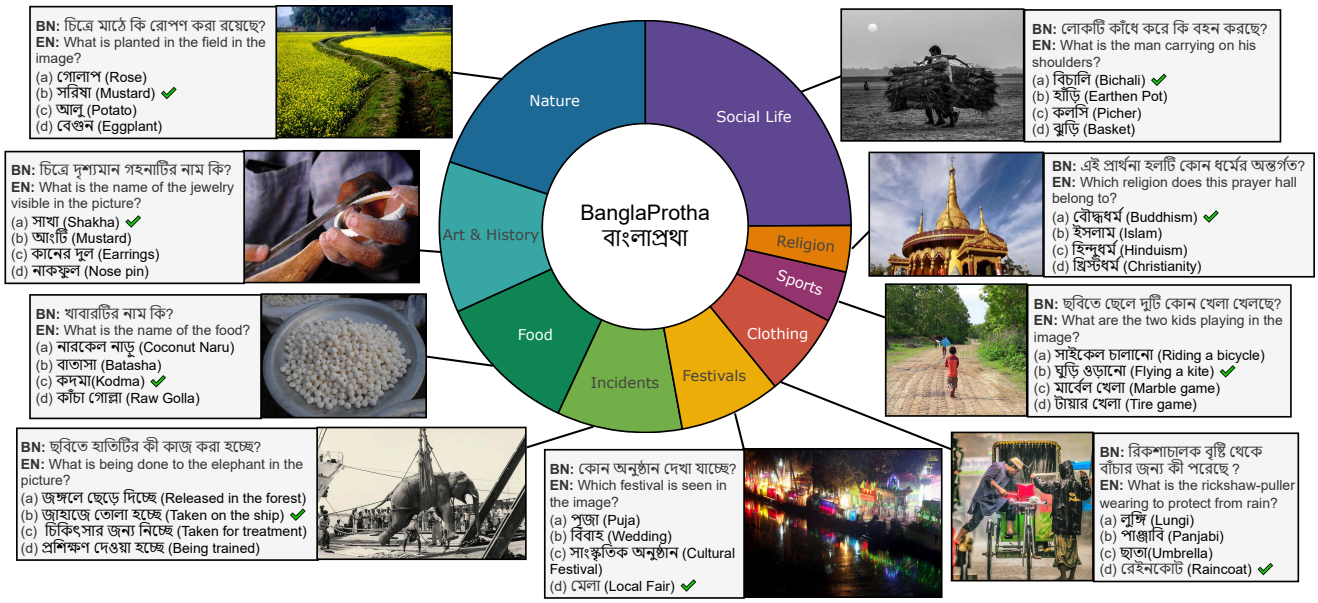

Figure 1. Sample images from the BANGLAPROTHA dataset across different cultural aspects, where (i) the images are relevant to Bengali cultural aspects, *i.e.*, event and fashion/attire (ii) the question is in native Bengali scripts with the English translation, and (iii) the answer options are semantically similar. The English translation has been provided for non-Bengali speakers and is *not* part of our dataset.

- Our findings reveal key behavioral insights on the cultural aspects, model size, necessity of training data & answer options, and vision-language alignment.

2. Related Works

Multilingual & Bengali VQA. While VQA research has predominantly been conducted in English [15], efforts have been made to develop non-English and multilingual VQA datasets, *e.g.*, FM-IQA [19], MCVQA [22], Multi30K [17], xGQA [49], and MaXM [15]. For Bengali, initial benchmarks like Bengali-VQA-v1 [26], derived from VQAv1 [7], and Bengali CLEVR [26], derived from CLEVR [28], were created via machine translation of English datasets. However, machine translation often struggles with low-resource languages like Bengali, as it fails to capture the nuances of the language and introduces linguistic artifacts [18, 40].

Rafi *et al.* [51] introduced a manually annotated Bengali VQA dataset derived from VQAv2 [21], but limited to binary questions and Western-centric images that fail to represent Bengali region-specific contexts. ChitroJera [8] and BVQA [13] addressed these issues using images relevant to the Bengali region. However, their QA pairs are generated using LLMs, which limits the questions' ability to reflect cultural nuances. Furthermore, both datasets serve as standard VQA benchmarks, without any categorization based on cultural concepts.

Western Bias & Multicultural VQA. Recent studies have highlighted the performance disparity across cultural and social norms, with VLMs exhibiting bias towards the Western counterparts [14, 44]. A lack of Bengali cultural understanding has also been evident [52]. Multicultural VQA benchmarks [11, 53] serve as a diagnostic tool in identifying such cultural biases, albeit with limited samples from each country, culture, or geographic region.

The CVQA dataset [53] evaluates VQA models on multi-lingual and multi-cultural contexts from 30 countries across 10 distinct cultural aspects. However, it simplifies the Bengali culture to the Indian region only, overlooking the diversity and traditions of the Bengali culture from Bangladesh. CultureVQA [45] establishes a similar benchmark across 11 countries, but categorized geographically. Thus, the Bengali culture was blended with the rest of the rich Indian culture. CultureVerse [32] provides the most diverse benchmark across 188 regions, but has limited samples in Bengali. Closest to our work is ALMBench [57], which includes Bengali cultural aspects from the whole region but lacks the nuances in answer options.

3. BANGLAPROTHA Dataset

Cultural Concepts Categorization. We systematically categorize BANGLAPROTHA into nine diverse categories encapsulating several Bengali cultural aspects as seen in Fig. 1. Our categorization draws inspiration from prior works in VQA [32, 39, 53], while adapting to better align with the Bengali cultural context. We aggregated several fine-grained categories into generalized ones: (1) vehicles & transportation, people, and everyday life to `Social`

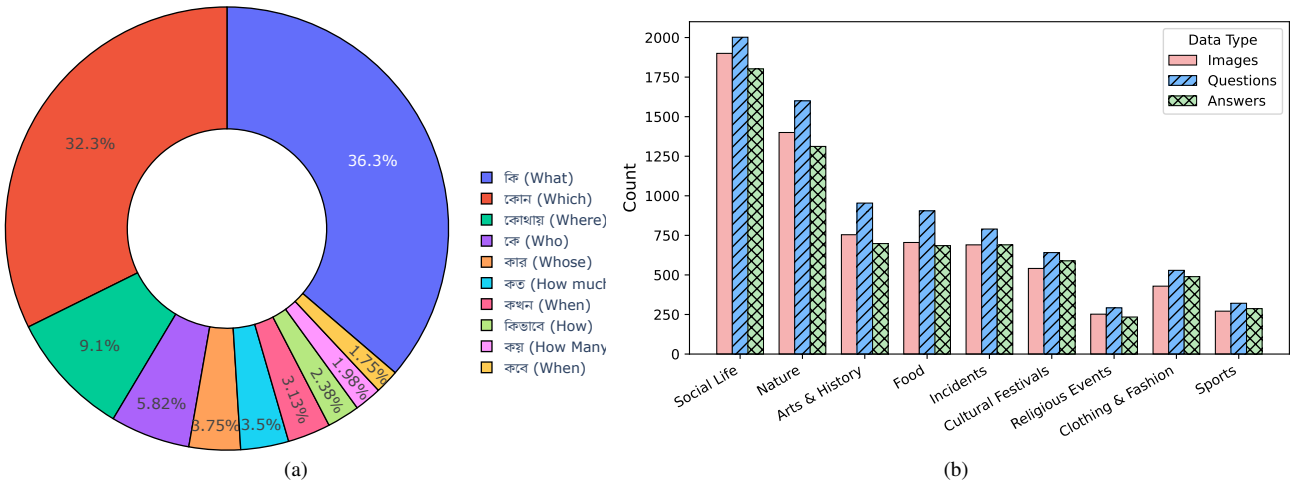

Figure 2. (a) Question type Distribution, (b) Category-wise count of images, questions, and unique answers across cultural concepts.

Life, (2) plants, animals, landscape, and buildings to Nature, (3) tradition, art, history, public figure, and pop-culture to Art and History. Detailed category descriptions are provided in §C.

Image Sourcing. We initially source 21,254 images from previous Bengali image datasets: BORNON [42], BANGLALEKHA-IMAGECAPTIONS [52], and BNATURE [4]. We further diversify the dataset by sourcing 2,800 publicly available Bengali images from websites. After rigorous filtering, the final distribution of image sources is provided in Tab. 1 and Fig. (§A.2a).

Data Annotation. We recruited 17 native Bengali annotators (12 male, 5 female), who are university undergraduates with strong expertise in Bengali culture and literature. The annotators received standard monetary compensation on a per-sample basis. To ensure annotation quality, we provided (i) detailed annotation guidelines (§A.1,A.2) and (ii) a dedicated annotation tool (§A.3).

For each image, the annotators created (i) a culturally relevant question and (ii) four semantically similar and question-relevant answer choices. They were instructed to carefully examine the image objects and their context to produce semantically similar answer options. For instance, if an image depicts a rural game, the annotators should include the names of other rural games as options. These semantically similar options were designed to test a model's ability to evaluate cultural elements. The annotators were also instructed to discard any image that did not belong to any of the nine culture categories.

Annotation Verification. Our dataset went through rigorous manual verification using a multi-stage filtering process. Samples were excluded based on several criteria: (1) images lacking sufficient cultural relevance, (2) imprecise or ambiguous questions, (3) misalignment between images and questions, (4) incorrect categorization, and (5) questions focusing solely on object detection without cultural context. From the initial 9,479 annotated image-question samples, 8,034 ($8034/9479 = 84.76\%$) met all quality standards and were retained, while the remaining 1,445 samples were discarded.

Data Source	#I	#Q	#Q:#I
BNATURE	2079	2517	1.21
BORNON	1402	1638	1.17
BANGLALEKHA	2541	2816	1.11
Bengali Websites	920	1063	1.16
Total	6942	8034	1.16

Table 1. Source distribution in BANGLAPROTHA. #I: No. of Images, #Q: No. of Questions, and #Q:#I: Question-Image ratio.

Dataset Statistics & Analysis. From Fig. 2b and Tab. (§A.3), we observe a higher sample count from the Social Life and Nature categories, whereas the count of unique images, questions, and answers remains somewhat uniform across the cultural aspects. "What" and "Which" type questions are predominantly more than other types, taking a substantial 68.6% of the total question types (Fig. 2a). Tab. 1 shows that BANGLAPROTHA has a strong representation of all four data sources while maintaining a good question-to-image ratio. Fig. (§A.2a) illustrates the distribution of cultural aspects across sources, *e.g.*, website images have a high number of Incidents, but fewer Religious Events. Finally, BANGLAPROTHA is divided into the standard 80:20 train-test splits.

Prompt Strategy	Prompt Overview
Zero-Shot/Base Prompt	*Find the most accurate option for the given image, question, and answer options.*
Chain-of-Thought (CoT)	*Think step by step before selecting the answer.*
Translation-based	*Translate the question and options into English.*
Culture-Specific	*The image is culturally relevant to the Bengali culture across nine key concepts.*
Description-based	*First describe the contents of the image and then use it to generate the answer.*

Table 2. The prompting strategies used in our benchmarks. The base prompt is appended at the end of the latter four prompting strategies.

4. Experiment Design

We classify our experiments into two categories: (i) prompt-based and (ii) fine-tuning experiments. The model details have been reported in (§D.1).

4.1. Prompt-based Experiments

We consider five prompting strategies (Tab. 2), the vanilla zero-shot prompting, Chain-of-Thought (CoT) [59], Translation-based [23], Culture-specific [32], and Description-based prompting. Description-based prompting takes inspiration from previous works where passing a textual description helped in visual classification [8, 32]. Tab. 2 provides an overview of all prompts, with detailed descriptions available in (§G).

4.2. Fine-tuning Experiments

Full Fine-tuning. To perform full fine-tuning on the pre-trained multimodal model ϕ_{mm}, we frame the task as a classification problem, following previous approaches [39, 54]. Given an image $\mathbf{I}$, a question $\mathbf{Q}$, and a set of answer choices $\{a_1, a_2, \ldots, a_n\}$, where n denotes the number of candidate answers, the model ϕ_{mm} is trained to predict the most relevant answer a_* from this set. We optimize the model using the standard **cross-entropy loss** over the answer choices, incentivizing the model to assign higher probabilities to the correct answers and vice versa.

Partial Fine-tuning. Most open-source VLMs adopt a modular Vision-Encoder + Adapter + LLM architecture. An input image $\mathbf{I}$ is first processed by a vision encoder $\phi_{\mathbf{I}}$, followed by an adapter module $\phi_{\mathbf{A}}$, which transforms image features into a sequence of visual tokens. In parallel, the input text $\mathbf{T}$ is encoded by a text embedding layer $\mathcal{E}(\cdot)$ to obtain textual tokens. These visual and textual tokens are concatenated and passed to a large language model ϕ_{LLM}, which performs multimodal reasoning and language generation. The output text prediction is given by:

$$\hat{\mathbf{T}} = \phi_{\text{LLM}}([\phi_{\mathbf{A}}(\phi_{\mathbf{I}}(\mathbf{I})); \mathcal{E}(\mathbf{T})]).$$

The entire model is trained using an autoregressive next-token prediction loss over the textual sequence $\mathbf{T}$. In our partial fine-tuning experiments, we explore three settings:

- *L-LoRA*: Applies LoRA [24] fine-tuning to the language model component ϕ_{LLM} within the multimodal model ϕ_{VLM}, aiming better alignment of the visual and textual representations.
- *L-LoRA + Adapter*: Fine-tunes both the adapter module ϕ_{A} and language model ϕ_{LLM} using LoRA, to investigate whether updating the visual token transformation improves performance.
- *L-LoRA + Adapter + VE*: Fine-tunes the vision encoder ϕ_{I} along with the previous components ϕ_{A} and ϕ_{LLM}, to examine whether learning fine-grained visual representations leads to better performance.

Training settings for both full and partial fine-tuning are provided in §D.2 and §D.3 respectively.

5. Results Analysis

The results of prompt-based experiments on eight open-source and three closed-source VLMs under five prompting strategies are presented in Tab. 3. Similarly, Tab. 4 reports the performance of five VLMs for full fine-tuning and another five for partial fine-tuning, each employing three distinct strategies as outlined in Sec. 4.2.

Open Source *vs*. Closed Source Models. The accuracy of monolingual open-source VLMs remains below 42% for both zero-shot and CoT prompting (Tab. 3). In contrast, multilingual open-source VLMs generally exceed 50% accuracy, except for the smaller Phi-3.5-V. The performance difference is obvious, as the monolingual models are not trained to comprehend Bengali text.

Among the open-source models, only Gemma-3 12B matches or exceeds the performance of the closed-source models. This can be attributed to the *supposedly* larger size or pretraining data of the closed-source models. GPT-4o consistently outperforms the other models, with the exception of being slightly surpassed by Gemma in zero-shot prompting and Claude in description-based prompting. Overall, GPT-4o takes the crown using the culture-specific prompting, achieving an average accuracy of 83.42%.

How should we prompt? From Fig. 3, we observe that closed-source and larger models, *e.g.*, llama-3.2-V 11B,

		Models	Cultural Concepts									
			Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
Zero-Shot Prompting	O-Mono	BLIP-2 OPT 6.7B [29]	35.21	33.38	31.47	43.77	32.36	46.13	38.42	34.18	40.66	37.29
		LLaVa-1.5 7B [30]	33.31	33.32	40.00	40.00	25.06	38.28	51.68	38.36	33.30	37.04
		LLaVa-Next 7B [31]	28.32	36.74	40.05	43.36	43.26	36.66	44.96	51.78	45.13	41.14
	O-Multi	LLaMa-3.2-V 11B [41]	61.23	69.78	60.52	56.30	72.67	73.34	70.20	70.05	83.32	68.60
		Phi-3.5-V [1]	38.31	31.72	23.28	25.08	29.94	45.05	36.71	30.10	38.34	33.17
		Phi-4 Multimodal [2]	57.64	47.53	44.14	42.37	53.56	54.42	53.27	50.88	50.20	50.45
		Qwen-2.5 7B [58]	50.10	50.15	60.20	51.66	73.26	60.08	61.72	73.32	80.05	62.28
		Gemma-3 12B [56]	68.28	86.73	71.72	76.74	86.72	78.26	83.36	83.28	78.37	**79.27**
	Closed	Claude-3.5 Sonnet [6]	81.67	70.24	81.90	71.92	75.08	82.26	81.38	72.12	82.56	77.68
		Gemini-2.0 Flash [20]	86.24	74.71	65.88	55.72	80.91	58.62	64.48	86.42	87.93	73.43
		GPT 4o [3]	78.68	68.52	78.93	76.74	75.35	91.32	83.41	72.28	79.38	78.29
Chain-of-Thought (CoT)	O-Mono	BLIP-2 OPT 6.7B	34.45	35.62	30.05	31.67	36.74	31.08	24.62	42.41	31.66	33.14
		LLaVa-1.5 7B	36.38	35.72	27.46	35.51	44.05	38.77	23.28	38.31	40.14	35.51
		LLaVa-Next 7B	43.08	38.27	42.44	38.31	30.53	28.80	60.10	25.84	46.57	39.33
	O-Multi	LLaMa-3.2-V 11B	65.05	68.31	58.27	63.32	76.26	78.28	80.04	65.03	76.27	70.09
		Phi-3.5-V	36.72	30.02	23.28	18.52	28.32	25.10	30.48	33.28	35.20	28.99
		Phi-4 Multimodal	47.49	48.22	49.94	23.24	56.64	48.28	46.72	48.22	44.92	45.96
		Qwen-2.5 7B	61.68	61.64	60.10	60.15	68.34	73.28	73.38	71.70	71.73	66.89
		Gemma-3 12B	75.04	88.28	73.34	75.06	90.04	75.02	83.25	78.32	81.69	80.00
	Closed	Claude-3.5 Sonnet	84.78	78.23	85.04	75.78	80.56	81.78	84.10	76.04	84.77	81.23
		Gemini-2.0 Flash	87.14	79.32	73.89	58.04	86.22	62.23	67.55	89.57	90.42	77.15
		GPT 4o	80.18	84.43	82.62	78.34	81.48	83.44	82.56	75.10	86.67	**81.65**
Translation-based	O-Mono	BLIP-2 OPT 6.7B	37.04	34.78	33.45	45.15	34.22	48.12	40.20	35.76	42.57	39.03
		LLaVa-1.5 7B	25.00	38.60	28.81	35.59	26.79	41.67	49.15	47.27	37.93	36.76
		LLaVa-Next 7B	33.33	35.00	37.29	31.67	25.42	36.67	38.33	31.67	33.33	33.63
	O-Multi	LLaMa-3.2-V 11B	62.71	75.00	66.10	65.00	74.58	63.33	70.00	63.33	85.00	69.45
		Phi-3.5-V	28.33	35.59	37.29	20.00	18.33	28.33	40.00	21.67	35.00	29.39
		Phi-4 Multimodal	55.17	45.00	45.76	42.37	56.67	51.67	48.33	60.00	51.67	50.74
		Qwen-2.5 7B	56.67	66.67	58.33	63.33	78.33	65.00	70.00	66.67	75.00	66.67
		Gemma-3 12B	71.67	85.00	75.00	76.27	86.67	76.67	83.33	85.00	85.00	80.52
	Closed	Claude-3.5 Sonnet	85.91	78.00	85.45	77.20	75.52	85.42	84.38	76.67	84.04	81.40
		Gemini-2.0 Flash	88.04	78.56	71.15	57.62	85.14	60.48	67.90	90.22	91.44	76.73
		GPT 4o	87.52	77.62	80.30	79.42	85.25	84.38	84.42	78.48	87.41	**82.75**
Culture-specific	O-Mono	BLIP-2 OPT 6.7B	33.92	32.29	30.14	42.64	31.32	44.63	37.22	32.90	39.66	36.08
		LLaVa-1.5 7B	45.00	38.33	30.51	27.12	30.00	23.73	35.59	36.67	40.00	34.11
		LLaVa-Next 7B	28.33	36.67	33.33	30.00	25.42	31.03	41.38	30.00	31.67	32.98
	O-Multi	LLaMa-3.2-V 11B	60.00	70.00	66.67	78.33	71.67	75.00	71.67	75.00	86.67	72.78
		Phi-3.5-V	37.93	40.68	31.48	24.14	32.76	38.98	35.71	29.82	27.59	33.23
		Phi-4 Multimodal	53.45	45.76	53.33	38.60	56.67	55.93	48.21	52.54	40.68	49.46
		Qwen-2.5 7B	58.62	66.10	55.93	56.90	72.88	61.67	69.49	70.00	80.00	65.73
		Gemma-3 12B	76.67	85.00	78.33	75.00	86.67	75.00	83.33	78.33	80.00	79.81
	Closed	Claude-3.5 Sonnet	86.62	79.23	87.32	77.67	78.14	87.11	85.78	78.20	86.15	82.91
		Gemini-2.0 Flash	90.52	80.48	74.60	58.78	87.04	63.67	70.42	91.05	92.88	78.83
		GPT 4o	81.92	86.71	84.23	80.24	84.05	83.10	85.30	75.94	89.25	**83.42**
Description-based	O-Mono	BLIP-2 OPT 6.7B	36.70	34.62	32.78	45.42	33.23	47.53	39.64	35.32	42.05	38.59
		LLaVa-1.5 7B	35.59	29.31	30.51	30.51	33.33	32.14	38.98	48.33	40.00	35.41
		LLaVa-Next 7B	31.67	42.37	31.03	29.31	28.57	36.67	37.29	40.68	50.00	36.40
	O-Multi	LLaMa-3.2-V 11B	62.71	77.59	66.10	67.80	79.66	66.10	71.43	70.69	83.05	71.68
		Phi-3.5-V	33.33	46.55	24.56	33.90	31.58	38.33	30.51	30.00	37.29	34.01
		Phi-4 Multimodal	54.39	44.07	55.93	37.93	56.67	50.00	46.67	50.00	49.15	49.42
		Qwen-2.5 7B	55.00	65.52	50.85	60.34	82.14	64.41	61.11	72.88	77.59	65.54
		Gemma-3 12B	70.00	86.67	75.00	76.67	83.33	75.00	83.33	78.33	80.00	78.70
	Closed	Claude-3.5 Sonnet	84.74	81.56	85.80	78.15	76.62	88.10	86.80	74.23	84.89	**82.32**
		Gemini-2.0 Flash	89.90	81.05	69.13	54.32	85.14	58.72	63.66	89.94	94.25	76.23
		GPT 4o	80.55	75.52	80.72	80.83	84.85	86.18	84.74	66.70	85.68	80.64

Table 3. Model benchmarking results across different ***Prompting Strategies*** on the test split of BANGLAPROTHA. *O-Mono* and *O-Multi* refer to the open-source monolingual and multilingual models, respectively, while *Closed* refers to the closed-source models. Cyan highlights the highest score for each cultural concept, and **Bold** indicates the overall best-performing model.

Models	Cultural Concepts									
	Food	Fest	Rel	Nature	Fash	Sport	Life	Art/Hs	Incid	Avg
Full Fine-Tuning										
BanglaBERT [12] + ViT [5]	33.24	33.27	37.18	36.08	31.36	38.14	40.17	33.82	32.21	35.05
BanglaBERT [12] + Swin [34]	35.45	36.25	38.78	37.20	32.45	40.17	41.27	35.92	33.64	36.79
CLIP [50]	31.25	32.90	36.78	34.64	28.62	33.15	34.70	32.34	30.48	32.76
LXMERT [55]	34.38	35.89	39.67	37.22	33.52	36.10	37.93	34.87	32.90	35.83
ALIGN [27]	38.91	39.83	41.62	40.63	35.44	43.39	44.72	39.60	35.84	40.00
SmolVLM2 [38]	32.34	48.09	42.56	41.65	45.76	41.87	41.00	52.48	55.89	44.63
Intern-VL3-2B [16]	65.12	40.24	45.08	53.36	45.13	61.60	56.58	56.53	46.67	**52.22**
Partial Fine-Tuning										
LLaVa-1.5 7B [30]										
L-LoRA	56.67	55.00	55.00	45.00	48.33	50.00	68.33	46.67	60.00	53.89
L-LoRA + Adapter	63.33	51.67	50.00	53.33	50.00	58.33	45.00	51.67	56.67	53.33
L-LoRA + Adapter + VE	66.67	58.33	36.67	56.67	53.33	60.00	60.00	63.33	73.33	**58.70**
Qwen2.5-VL 7B [58]										
L-LoRA	60.00	76.67	63.33	66.67	78.33	76.67	76.67	78.33	76.67	72.59
L-LoRA + Adapter	68.33	76.67	61.67	66.67	71.67	78.33	70.00	80.00	83.33	72.96
L-LoRA + Adapter + VE	65.00	70.00	63.33	75.00	71.67	81.67	80.00	85.00	78.33	**74.44**
Paligemma-2 10B [10]										
L-LoRA	73.33	78.33	66.67	66.67	86.67	75.00	70.00	70.00	68.33	72.78
L-LoRA + Adapter	81.67	76.67	63.33	80.00	81.67	76.67	70.00	81.67	76.67	**76.48**
L-LoRA + Adapter + VE	68.33	83.33	70.00	65.00	81.67	73.33	66.67	76.67	78.33	73.70
LLaMa-3.2V 11B [41]										
L-LoRA	63.33	76.67	61.67	63.33	75.00	71.67	73.33	71.67	76.67	70.37
L-LoRA + Adapter	81.67	70.00	68.33	60.00	78.33	71.67	75.00	73.33	81.67	**73.33**
L-LoRA + Adapter + VE	56.67	71.67	61.67	66.67	70.00	66.67	65.00	75.00	76.67	67.78
Gemma-3 12B [56]										
L-LoRA	80.00	81.67	75.00	68.33	86.67	68.33	71.67	81.67	83.33	77.41
L-LoRA + Adapter	81.67	83.33	70.00	71.67	85.00	78.33	73.33	83.33	76.67	78.15
L-LoRA + Adapter + VE	82.22	92.12	80.64	81.90	88.45	84.52	75.72	85.13	86.44	**84.13**

Table 4. Model benchmarking results across different ***Finetuning Strategies*** on the test split of BANGLAPROTHA. *O-Mono* and *O-Multi* refer to the open-source monolingual and multilingual models, respectively, while *Closed* refers to the closed-source models. Cyan highlights the highest score for each cultural concept, and **Bold** indicates the overall best-performing model in each category.

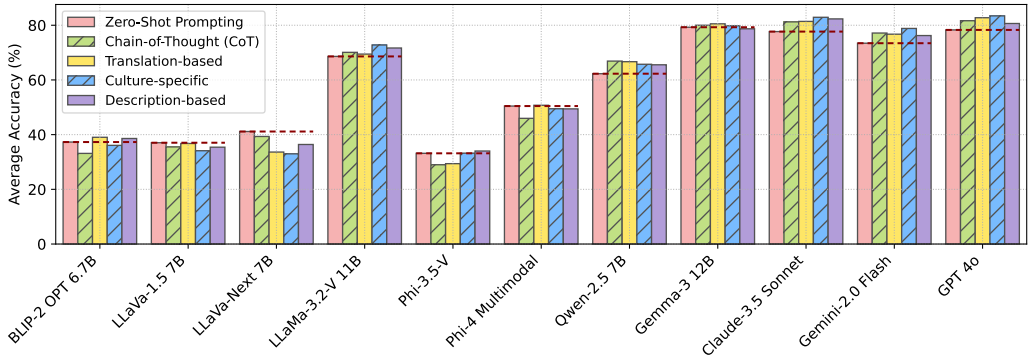

Figure 3. Average accuracy of models across prompting strategies on the test split of BANGLAPROTHA.

Gemma-3 12B, and Qwen 2.5 7B, achieve the highest performance gains under culture-specific prompting, emphasizing the need for culturally tailored instructions for reasoning and generating culturally relevant responses. Zero-shot prompting, however, remains the weakest of the prompting strategies. In sharp contrast, zero-shot prompting consistently outperforms the other strategies for the smaller open-source models, suggesting that the additional prompting instructions tend to overwhelm the limited capacity of the smaller architectures.

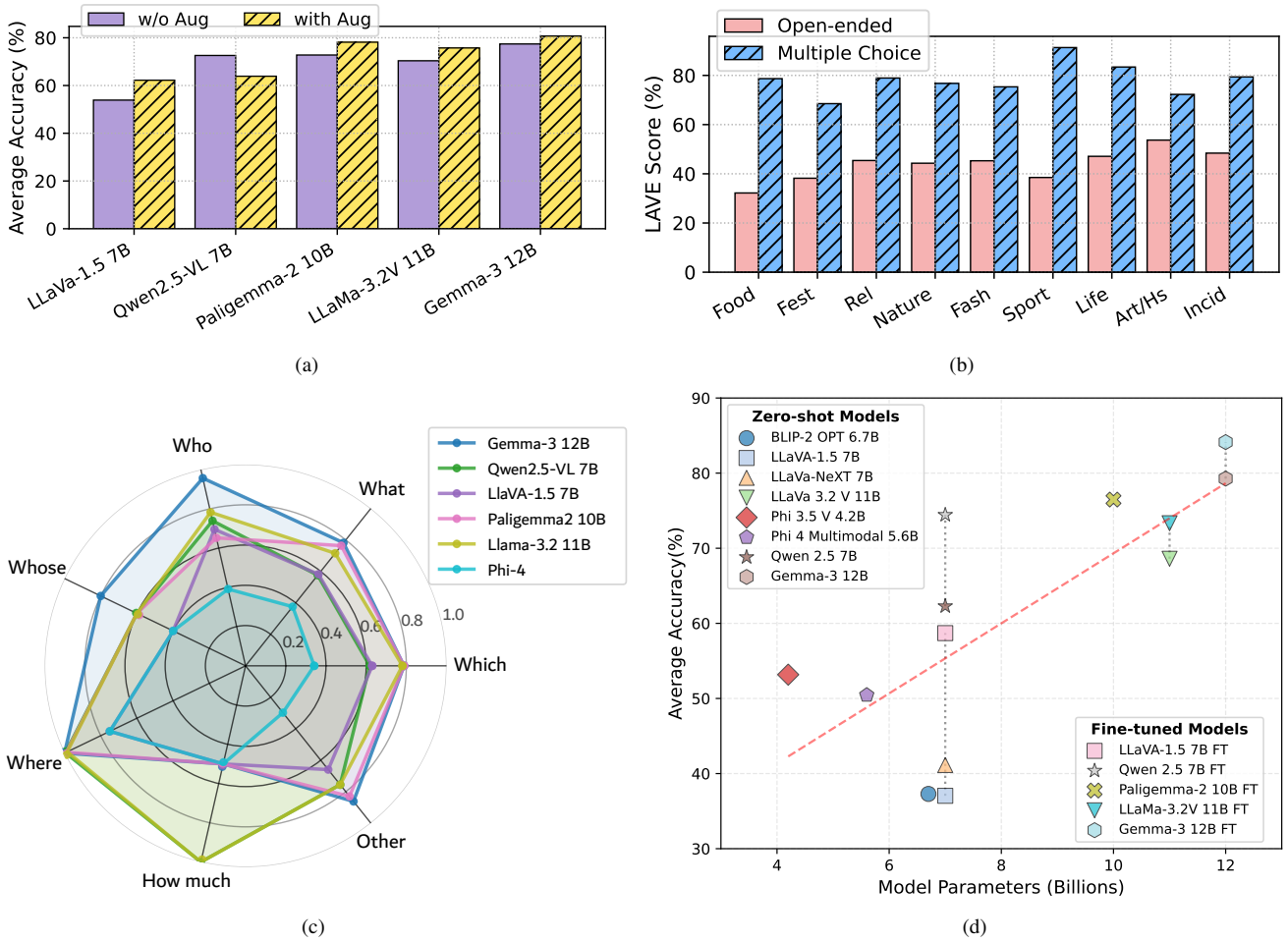

Figure 4. (a) L-LoRA accuracy of models with *vs*. without data augmentation, (b) LAVE scores of open-ended *vs*. multiple-choice questions across cultural concepts for GPT-4o, (c) Accuracy of fine-tuned models across question types, and (d) Average accuracy *vs*. model parameters of zero-shot and fine-tuned models (using the best-performing strategy), with the accuracy trendline across parameters (in red).

Prompting *vs*. Fine-tuning. Following Tabs. 3 and 4, fully fine-tuned dual encoder and modality alignment models achieved performance comparable to the monolingual open-source models (∼40% average accuracy). Partial fine-tuning showed a substantial bump in performance over prompting strategies, *e.g*., the best fine-tuning strategy for LLaVa-1.5 7B outperformed its best prompting strategy by 21.66%. Similarly, fine-tuned Gemma-3 12B achieved the highest average accuracy, 84.13%, on our dataset, slightly surpassing GPT-4o using culture-specific prompting.

What's the best way to finetune? Tab. 4 highlights ALIGN outperforming other fully fine-tuned models across all aspects, but lagging behind the larger partially fine-tuned LLMs. Fine-tuning the language component using L-LoRA usually improved the performance, *e.g*., for LLaVa-1.5 by roughly 17% *vs*. its best prompting strategy. However, we observed several instances of performance drop, *e.g*., Gemma-3's average accuracy dropped roughly 3% *vs*. its best prompting strategy.

Should we fine-tune the visual component? Fine-tuning the visual components, *i.e*., adapters and vision encoders, generally leads to improved performance (Tab. 4). The best results are usually achieved when *both*, visual and textual, components are fine-tuned, improving the alignment with the underrepresented Bengali cultural images. However, models, such as Paligemma-2 10B and LLaMa-3.2V 11B, underperform when adapters and vision encoders are fine-tuned, likely due to overfitting. Adapter fine-tuning tends to outperform applying L-LoRA fine-tuning only, though occasionally leading to a marginal performance decline (*e.g*., -0.65% accuracy drop for LLaVa-1.5 7B).

Performance variation across Cultural Aspects. Most models tend to perform better on `Fashion`, `Social`

(a) (b) (c)

Figure 5. Performance comparison across evaluation methods of Gemma-3 12B: (a) unanimous correct predictions, (b) only LoRA-FT predicts correctly, and (c) unanimous incorrect predictions. Red and Green highlight the incorrect and correct predictions respectively.

`Life`, and `Incidents` categories, while struggling with `Religious` and `Nature` Fig. (§A.3b). Under zero-shot setting, GPT-4o excels in `Sports`, Gemini-2.0 in `Art/Hs`, `Incidents`, and `Food`, Gemma-3 in `Fashion` and `Festivals`, and Claude-3.5 in `Religion`. We see similar performance strengths of Gemini-2.0 and Gemma-3 on `Festival`, `Fashion`, and `Art/Hs`, likely due to overlap in training corpora as the models share the same parent company. Gemma-3, using L-LoRA + Adapter + VE, consistently outperforms most fine-tuned models across all cultural aspects.

Performance across Question Types. Following Fig. 4c, the fine-tuned Llama-3.2 11B performs better on *How much* questions, while Gemma-3 12B leads on rest of the question types. Similarly, Fig. (§A.3a) shows the best prompting model, GPT-4o, performing better on *How much* (93.4%), *When* (90.6%), and *Where* (87.2%) questions, demonstrating the model's excellence on qualitative, temporal, and spatial reasoning. In contrast, the lowest performance was observed for *How Many* (66.5%) questions, exposing the model's weakness in counting.

Impact of Data Augmentation. We applied LLaVA-style augmentation [30] during L-LoRA finetuning by augmenting each question-answer(QA) pair $k = 4$ times, k representing the number of options per question, resulting in $6.5k \times 4 = 26k$ samples (detailed in §D.4). This compensates for the scarcity of multiple-choice variations by shuffling the answer options to create new training samples. From Fig. 4a, we observe a 3-9% boost in average accuracy across all models except Qwen-2.5 (cultural aspect-wise breakdown in Tab. §A.4). As the augmentation only permutes the position of the correct option, the results expose a positional textual bias in the models. Similar experiments were conducted using the circular evaluation strategy [33], reported in §E.3 and Tab. A.8.

What if we remove answer options? We investigate this by evaluating GPT-4o in an open-ended setting using the LAVE metric [36] (details in §E.1). In this setup, GPT-4o attained a LAVE score of 43.68%, a significant drop from its 81.64% accuracy in the multiple-choice format. Fig. 4b shows performance dropping notably for `Food`, `Festival`, and `Sport` categories. Similar experiments were conducted on open-source VLMs (Tab. A.6), with models exhibiting a consistent decline in performance under the open-ended setting.

Error Analysis. Fig. 5 shows predictions of the best performing Gemma-3 12B model across prompting and fine-tuning strategies. In Fig. 5a, all evaluation settings correctly identify the traditional attire (*Panjabi*). In Fig. 5b, only the LoRA-FT setting correctly recognizes the correct habitat of the animal. In Fig. 5c, all settings mispredict the location of the floating guava market, which requires a high level of cultural knowledge even for Bengali natives. While some classes of culturally-grounded visual questions can be handled by the fine-tuned model, others that require in-depth knowledge remain challenging. §F expands qualitative error analysis across models and cultural concepts.

6. Conclusion

We introduced BANGLAPROTHA, the first Bengali culturally grounded Visual Question Answering dataset encompassing nine diverse cultural domains. Through comprehensive experiments with both open- and closed-source, monolingual and multilingual VLMs under various fine-tuning and prompting setups, we provide valuable insights into the current capabilities and limitations of Bengali multimodal understanding. We hope our dataset and findings will foster future research toward more culturally aware and linguistically inclusive vision-language models.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 5, 3, 8

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 5, 3, 8

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 3

[4] Hasan Al Faraby, Md Muzahidul Azad, Md Riduyan Fedous, Md Kishor Morol, et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE, 2020. 3

[5] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 6, 3

[6] Anthropic. Model card addendum for Claude 3. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf, 2024. Accessed: 2025-06-04. 5, 3

[7] 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, 2

[8] Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, Fabiha Haider, Fariha Tanjim Shifat, Md Tasmim Rahman Adib, Anam Borhan Uddin, Md Farhan Ishmam, and Md Farhad Alam. Chitrojera: A regionally relevant visual question answering dataset for bangla. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 473–491. Springer, 2025. 2, 4

[9] Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. Viscounth: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20, 2023. 1

[10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1, 6, 3

[11] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024. 2

[12] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics. 6, 3

[13] Md Shalha Mucha Bhuyan, Eftekhar Hossain, Khaleda Akhter Sathi, Md Azad Hossain, and M Ali Akber Dewan. Bvqa: Connecting language and vision through multimodal attention for open-ended question answering. *IEEE Access*, 2025. 2

[14] 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

[15] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*, 2022. 2

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 6

[17] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 2

[18] Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265, 2024. 2

[19] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. 2

[20] Google DeepMind and Sundar Pichai. Introducing Gemini 2.0: Our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, December 2024. Accessed: 2025-06-04. 5, 3

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[22] Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 900–913, 2020. 2

[23] Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua,

Md Fahim, and Md Farhad Alam Bhuiyan. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. 4

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1, 4, 3

[25] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270, 2024. 1

[26] SM Shahriar Islam, Riyad Ahsan Auntor, Minhajul Islam, Mohammad Yousuf Hossain Anik, ABM Alim Al Islam, and Jannatun Noor. Note: Towards devising an efficient vqa in the bengali language. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 632–637, 2022. 2

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 6, 3

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 2

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5, 3

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5, 6, 8, 3, 9

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 5, 3

[32] 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, 4

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 8, 6

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 3

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[36] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024. 8

[37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[38] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 6

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3190–3199, 06 2019. 2, 4

[40] Raphael Merx, Adérito José Guterres Correia, Hanna Suominen, and Ekaterina Vylomova. Low-resource machine translation: what for? who for? an observational study on a dedicated tetun language translation service. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jonathan Washington, Nathaniel Oco, and Xiaobing Zhao, editors, *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 54–65, Albuquerque, New Mexico, U.S.A., May 2025. Association for Computational Linguistics. 2

[41] Meta AI. Llama 3: Connecting the next generation of ai with vision, edge, and mobile devices at Connect 2024. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, May 2024. Accessed: 2025-06-04. 5, 6, 3, 8, 9

[42] Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. Bornon: Bengali image captioning with transformer-based deep learning approach. *SN Computer Science*, 3:1–16, 2022. 3

[43] Ghulam Murshid. *Bengali culture over a thousand years*. Niyogi Books, 2018. 1

[44] 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

[45] 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, 2

[46] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. *arXiv preprint arXiv:2410.17250*, 2024. 1

[47] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024. 1

[48] ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *arXiv preprint arXiv:2406.16469*, 2024. 5

[49] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*, 2021. 2

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 6, 3

[51] Mahamudul Hasan Rafi, Shifat Islam, SM Hasan Imtiaz Labib, SM Sajid Hasan, Faisal Muhammad Shah, and Sifat Ahmed. A deep learning-based bengali visual question answering system. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 114–119. IEEE, 2022. 2

[52] Matiur Rahman, Nabeel Mohammed, Nafees Mansoor, and Sifat Momen. Chittron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642, 2019. 1, 2, 3

[53] 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, 2

[54] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 4

[55] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 6, 3

[56] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 5, 6, 3, 8, 9

[57] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575, 2025. 2

[58] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5, 6, 3, 8, 9

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4

[60] Thomas Wolf, Lysandre Debut, Victor Sanphih, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 3

[61] Pengju Xu, Yan Wang, Shuyuan Zhang, Xuan Zhou, Xin Li, Yue Yuan, Fengzhao Li, Shunyuan Zhou, Xingyu Wang, Yi Zhang, et al. Tcc-bench: Benchmarking the traditional chinese culture understanding capabilities of mllms. *arXiv preprint arXiv:2505.11275*, 2025. 1

[62] 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. 3