Many Dialects, Many Languages, One Cultural Lens: Evaluating Multilingual VLMs for Bengali Culture Understanding Across Historically Linked Languages and Regional Dialects

Anonymous ACL submission

Abstract

Bangla culture is richly expressed through region, dialect, history, food, politics, media, and everyday visual life, yet it remains underrepresented in multimodal evaluation. To address this gap, we introduce **BanglaVerse**, a culturally grounded benchmark for evaluating multilingual vision–language models (VLMs) on Bengali culture across historically linked languages and regional dialects. Built from 1,152 manually curated images across nine domains, the benchmark supports visual question answering and captioning, and is expanded into four languages and five Bangla dialects, yielding ∼32.3K artifacts. Our experiments show that evaluating only standard Bangla overestimates true model capability: performance drops under dialectal variation, especially for caption generation, while historically linked languages such as Hindi and Urdu retain some cultural meaning but remain weaker for structured reasoning. Across domains, the main bottleneck is missing cultural knowledge rather than visual grounding alone, with knowledge-intensive categories. These findings position BanglaVerse as a more realistic test bed for measuring culturally grounded multimodal understanding under linguistic variation.[1]

1 Introduction

Multilingual vision–language models (MVLMs) are increasingly deployed in culturally and linguistically diverse settings, yet their evaluation remains overwhelmingly centered on standard language varieties (Nayak et al., 2024). For Bangla, existing multimodal resources are still scarce and limited in scope (Pawar et al., 2025). Most available datasets focus narrowly on visual question answering and often depend heavily on synthetic annotation pipelines (Barua et al., 2024), resulting in benchmarks that insufficiently capture culturally grounded and linguistically authentic usage. Recent work on culturally specific benchmarks in other regions has further shown that multimodal evaluation must go beyond surface-level translation and account for cultural context, local knowledge, and linguistic diversity (Faraz et al., 2025; Liu et al., 2025). However, there remains no comprehensive dialect-aware benchmark for evaluating MVLMs on Bangla culture.

A central limitation of existing evaluation practices is that they treat standard Bangla as a sufficient proxy for Bengali cultural understanding. In reality, culture is not expressed only through standardized language, but also through regional dialects, historically linked languages, and locally situated forms of description, reference, and reasoning (Adilazuarda et al., 2024). As a result, evaluating only standard Bangla underestimates how fragile multilingual VLMs can be when the same cultural content is expressed through dialectal and cross-lingual variation. A model that appears competent in standard Bangla may still fail to preserve cultural meaning, visual grounding, or linguistic consistency when tested on regional varieties or closely related other languages.

To address this gap, we present BanglaVerse, a culturally grounded multilingual and multidialectal benchmark for evaluating MVLMs on Bengali culture across standard Bangla, historically linked languages, and regional dialects. BanglaVerse is built from a manually curated, image-centric core dataset spanning nine culturally rich domains, and is extended across four languages: Bangla, English, Hindi, and Urdu; as well as five Bangla dialects: Barishal, Chittagong, Noakhali, Rangpur, and Sylhet. The benchmark supports two image-grounded tasks: Visual Question Answering (VQA) and Image Captioning (CAP). By design, BanglaVerse enables a more realistic and fine-grained study of cultural understanding under linguistic variation, allowing us to analyze not only model performance, but also robustness, transfer,

[1] anonymous.4open.science/BanglaVerse

Dataset	VQA	CAP.	#ART.	CA	CMA	ML	DIA	Ann.
Bengali VQA (Hasan et al., 2025)	✓	✗	5,000	✗	✗	✗	✗	Google Trans.
Bengali CLEVR (Hasan et al., 2023)	✓	✗	12,291	✗	✗	✗	✗	Google Trans.
Bengali VQA 2.0 (Rafi et al., 2022)	✓	✗	13,046	✗	✗	✗	✗	Manual
CVQA (Romero et al., 2024)	✓	✗	286	✗	✓	✗	✗	Manual
ChitroJera (Barua et al., 2024)	✓	✗	15,292	✓	✓	✗	✗	GPT-4 Turbo
BanglaProtha (Fahim et al., 2026)	✓	✗	8,034	✓	✓	✗	✗	Manual
BanglaVerse (Ours)	✓	✓	**32,256**	✓	✓	✓	✓	Manual+†

Table 1: Comparison with related Bangla and culturally grounded multimodal benchmarks. VQA = Visual Question Answering, CAP. = Captioning, #ART. = total number of artifacts, CA = Cultural Awareness, CMA = Categorical Metadata Availability, ML = Multilingual, DIA = Dialectal coverage, Ann. = Annotation strategy, and Google Trans. = Google Translate. †BanglaVerse is manually curated at the source level and later expanded into multiple languages and dialects through controlled automated translation and quality-checking pipelines.

and failure patterns across languages and dialects. Our main contributions are as follows:

- We introduce **BanglaVerse**, a culturally grounded multilingual and multidialectal benchmark for Bengali culture understanding, built on 1,152 manually curated images with high-quality annotations verified through a collaborative cross-checking process.
- We expand the benchmark to four languages and five Bangla dialects, yielding ∼32.3K total artifacts across captioning and visual question answering tasks, and enabling systematic evaluation across historically linked languages and regional varieties.
- We conduct a comprehensive empirical study of state-of-the-art multilingual VLMs under multiple prompting strategies, showing that evaluation only on standard Bangla can obscure substantial weaknesses that emerge under dialectal and cross-lingual variation.

2 Background and Related Works

2.1 Background

Bangla has evolved through a long history of cultural, political, and linguistic exchange across South Asia, making Bengali cultural expression deeply connected to neighboring languages and regional speech varieties (Shahen et al., 2019). Motivated by this history, we extend our benchmark beyond standard Bangla to English, Hindi, and Urdu, which remain important due to their historical, cultural, and communicative links with Bangla, as well as their relevance in education, media, and digital communication (Van Schendel, 2020). At the same time, evaluating Bengali culture only through standard Bangla is insufficient, since much of Bangladesh's cultural life is expressed through regional dialects and locally grounded forms of speech (Hasan and Rahaman, 2014; Karmaker, 2019). We therefore include five major Bangla dialects, Barishal, Chittagong, Noakhali, Rangpur, and Sylhet, which together capture a broad range of socially and culturally salient variation across Bangladesh.

2.2 Related Works

A growing body of work has emphasized that multimodal evaluation should reflect cultural context rather than relying solely on generic or Western-centric imagery. In Bangla, several multimodal datasets have been introduced, but most remain limited either in cultural scope or annotation authenticity. Bengali VQA (Hasan et al., 2025), Bengali CLEVR (Hasan et al., 2023), and Bengali VQA 2.0 (Rafi et al., 2022) primarily target visual question answering, but the first two are translated from English benchmarks while the third remains relatively narrow in cultural coverage. CVQA (Romero et al., 2024) and ChitroJera (Barua et al., 2024) move closer to culturally aware evaluation, yet CVQA is small in scale and ChitroJera relies heavily on GPT-4 Turbo-generated annotations, which may limit authenticity and grounding in real cultural usage.

A recent and closely related effort is BanglaProtha (Fahim et al., 2026), which studies Bengali cultural understanding as a long-tail multimodal evaluation problem. It introduces a Bengali cultural VQA benchmark organized around multiple cultural aspects, with native Bengali questions and semantically similar answer options, and evaluates different classes of VLMs under prompt-

ing and fine-tuning settings. While it's a vital step forward for representing Bengali culture, it strictly focuses on VQA and misses the region's rich dialectal and multilingual variations. Beyond Bangla, there is growing interest in culturally grounded multimodal benchmarks for other languages and regions. Maji et al. (2025) introduced DRISHTIKON, a multilingual benchmark centered on Indian cultural contexts, while Alwajih et al. (2025) proposed PALM, a large-scale, community-driven dataset spanning all 22 Arab countries with instruction–response pairs in both Modern Standard Arabic and dialectal variants. These efforts highlight the importance of cultural inclusivity, contextual grounding, and broader linguistic coverage in multimodal and language model evaluation.

Despite this progress, an important limitation remains across existing Bangla benchmarks: evaluation is almost always conducted only in standard Bangla. In practice, however, culture is expressed not only through standardized language but also through regional dialects, local lexical choices, and historically linked languages (Keleg, 2025). As a result, existing resources cannot reveal whether multilingual VLMs genuinely understand Bengali culture or merely perform well on standardized forms of it. A comparison of existing Bangla and culturally grounded multimodal benchmarks is provided in Table 1.

3 Dataset Creation

In this section, we present the overall pipeline for the BanglaVerse dataset, shown in Figure 1.

3.1 Data Collection

We collected images and associated textual context from a combination of online and offline sources, including news articles,[2] Wikipedia,[3] Banglapedia,[4] historical books,[5] and general knowledge books.[6] We organized the collected materials into nine major domains: *Culture, Food, History, Media & Movies, National Achievements, Nature, Personalities, Politics, and Sports*. These specific domains were carefully selected because they collectively encapsulate the multifaceted nature of Bengali identity. Rather than defining culture strictly through traditional customs or attire, we adopt a holistic framework where historical milestones, regional geography, traditions, public figures, and everyday entertainment all contribute to the societal ethos. Each domain contains an `images` directory and an `annotations` directory.

3.2 Annotation Setup

The dataset was annotated by five authors through a multi-stage collaborative process. Four authors conducted the primary annotation, while a fifth author acted as an adjudicator to resolve disagreements and maintain consistency across domains. Each annotator was initially assigned two domains, with one annotator covering three domains to account for all nine domains. After data collection from heterogeneous sources, the annotated instances were randomized and reassigned for cross-verification. This process helped minimize annotator-specific bias, improve consistency, and ensure that all textual outputs remained grounded in the associated images.

We followed a shared annotation protocol throughout the process. Annotators were instructed to ensure that all captions and questions were directly supported by visual evidence, culturally appropriate, and linguistically natural in Bangla. For VQA, they were additionally asked to construct plausible distractors while maintaining a single unambiguous correct answer. For captioning, the goal was to produce concise yet informative descriptions that captured both the visible content and its culturally relevant context.

Disagreements most commonly involved the degree of cultural specificity to include, the interpretation of culturally nuanced visual elements, and the formulation of semantically similar but clearly incorrect answer options. Such cases were discussed among the annotators and, where necessary, resolved by the adjudicating author. To assess reliability, we computed inter-annotator agreement on a cross-verified subset using Cohen's kappa and obtained a score of $\kappa = 0.87$, indicating strong agreement.

3.3 Artifacts Generation

We generated the final benchmark artifacts through a two-stage pipeline built on top of the manually curated Bangla source dataset. Each source instance consists of an image paired with two image-grounded task annotations: a caption and a VQA item. The source Bangla annotations served as the

[2] https://allbanglanewspapersbd.com/
[3] https://www.wikipedia.org/
[4] https://en.banglapedia.org/
[5] goodreads/history-bangladesh
[6] rokomari/general-knowledge

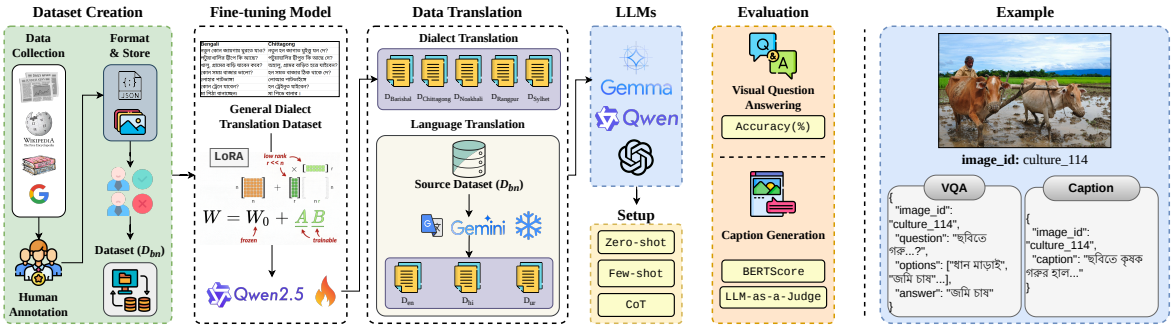

Figure 1: Overview of the BanglaVerse dataset and experimental setup. The figure shows the two task types, example annotations for each task, artifacts generation and evaluation pipeline with multiple metrics.

canonical reference point for all subsequent multilingual and dialectal artifact generation.

In the first stage, we performed multilingual expansion by translating the Bangla annotations into three historically linked languages: English, Hindi, and Urdu. For each target language, we used `Gemini-2.5-Flash` to translate both captions and VQA instances. During this process, we aimed to preserve not only the literal semantic content of the source annotation but also its cultural specificity, pragmatic intent, and alignment with the visual content. In particular, we took care to ensure that culturally salient references, named entities, and locally meaningful expressions were retained as faithfully as possible rather than being replaced with more generic alternatives.

In the second stage, we generated multidialectal variants of the benchmark to capture regional linguistic diversity within Bangla itself. We considered five major Bangla dialects: Barishal, Chittagong, Noakhali, Rangpur, and Sylhet. Rather than directly prompting a general-purpose model to produce dialectal outputs, we adopted a dedicated dialect generation strategy. We first fine-tuned the `Qwen2.5-3B-Instruct` model on a native standard-Bangla-to-dialect translation dataset called BanglaDial (Mahi et al., 2025), so that it could better capture lexical, morphological, and stylistic features specific to each dialect. See Appendix A for additional details regarding the fine-tuning model. The fine-tuned model was then used to convert the source Bangla captions and VQA items into dialect-specific forms. By keeping the image fixed and only varying the linguistic form, we created a controlled setting for studying how model performance changes when the same Bengali cultural content is expressed through different dialectal varieties.

To refine artifact quality, we applied an automated post-generation verification stage using `Gemini-2.5-Flash`. This step evaluated all multilingual and dialectal outputs for semantic fidelity, fluency, and cultural preservation, allowing us to identify and manually revise instances that were grammatically acceptable but culturally weakened.

To rigorously validate the authenticity of our generated artifacts, we also conducted a comprehensive human evaluation on a representative subset of the data. Five senior Computer Science (CS) undergraduates familiar with Large Language Models (LLMs), each a native speaker of one of the five different dialects, evaluated the dialectal outputs. Concurrently, three of the authors performed back-translation validation on the multilingual variants. In total, 50 randomly sampled instances per language and dialect were manually assessed. Evaluators assigned scores using a discrete 3-point scale: 0 (Inaccurate / Unusable), 1 (Acceptable with Minor Revisions), and 2 (Flawless / Highly Authentic). The results demonstrated exceptional generation quality and authenticity, with 94% of the evaluated artifacts receiving the maximum score of 2. Further details regarding the human evaluation are available in Appendix B.

3.4 Dataset Statistics

Table 2 summarizes the overall corpus statistics of BanglaVerse across nine cultural domains, four languages, and five Bangla dialects. The benchmark comprises a foundational set of 1,152 unique images. To evaluate multidialectal and multilingual capabilities, the base annotations for each image were expanded across nine distinct linguistic variations. Consequently, the base caption and VQA counts are multiplied by 9, resulting

Type	Cult.	Food	Hist.	M&M	Nat. Achv.	Nature	Pers.	Pol.	Sports	Total
Images	114	150	150	150	75	150	150	150	63	1,152
Captions	9×114	9×150	9×150	9×150	9×75	9×150	9×150	9×150	9×64	10,368
VQA	9×228	9×301	9×308	9×288	9×148	9×299	9×300	9×306	9×125	20,727
Total	**3,192**	**4,209**	**4,272**	**4,092**	**2,082**	**4,191**	**4,200**	**4,254**	**1,764**	**32,256**

Table 2: Overall corpus statistics across 4 languages and 5 Bangla dialects. The 9× notation illustrates the 9 linguistic variants. Abbreviations: Cult. = Culture, Hist. = History, M&M = Media & Movies, Nat. Achv. = National Achievements, Pers. = Personalities, Pol. = Politics, and VQA = Visual Question Answering.

in 10,368 captions and 20,727 VQA pairs, yielding 32,256 artifacts overall. Domains such as *Food*, *History*, *Media & Movies*, *Nature*, *Personalities*, and *Politics* contribute the largest number of samples, whereas *National Achievements* and *Sports* remain comparatively smaller. Among all domains, *History* and *Politics* contain the highest number of VQA annotations. In addition to scale, the benchmark contains linguistically rich annotations. The average caption length is approximately 90–140 characters, while the average VQA question length ranges from 50–80 characters. This indicates that the dataset goes beyond very short or template-like annotations and supports more meaningful evaluation of culturally grounded multimodal understanding.

3.5 Overview of Tasks

Visual Question Answering. For each image, we construct a VQA instance designed to test both direct visual understanding and culturally grounded commonsense reasoning. Each question is formulated based on the visible image content, while also requiring the model to interpret culturally specific practices, objects, events, or implicit context when necessary. To maintain consistency and evaluability, each VQA item is paired with four multiple-choice answer options, consisting of one correct ground-truth answer and three carefully crafted distractors. For example, for `image_id: culture_114`, we include the question: ছবিতে গরু ব্যবহার করে কী কাজ করা হচ্ছে? *(What task is being performed using the cow in the image?)* with answer options ["ধান মাড়াই","জমি চাষ","শস্য মজুদ","গো-খাদ্য প্রস্তুত"] *(["Threshing rice", "Plowing the field", "Storing crops", "Preparing cattle feed"])*, where the correct answer is "জমি চাষ" *("Plowing the field")*. To ensure rigorous evaluation and prevent models from relying on superficial language biases or simple process-of-elimination, the distractors are designed to be culturally relevant and visually plausible. Instead of using random or out-of-context phrases, we curate incorrect options that represent related regional activities, similar objects, or plausible alternative scenarios (e.g., "Threshing rice" is a valid agricultural task but incorrect for the specific image).

Caption Generation. Each image is also paired with a natural language caption describing the visual content in Bengali. The captions are designed to provide concise yet culturally informative descriptions of the entities, actions, and contexts depicted in the image. Rather than merely listing visible objects, the captions often include culturally meaningful details that help situate the image within Bengali social, historical, or everyday life. For instance, for `image_id: culture_114`, the caption is: "ছবিতে কৃষক গরুর হাল দিয়ে জমি চাষ করছেন, যা গ্রামীণ কৃষির ঐতিহ্য।" *(The farmer is plowing the field with oxen, a tradition of rural agriculture.)*. This task evaluates whether a model can generate fluent and visually grounded descriptions while preserving the cultural significance.

4 Experimental Setup

4.1 Models

We evaluate a diverse set of multilingual and multimodal models spanning both open-source and proprietary systems. Our experiments include `Gemma3-4B`, `Gemma3-12B`, `Gemma3-27B` (Kamath et al., 2025), `Qwen2.5-VL-7B` (Qwen et al., 2024), `Qwen3-VL-8B` (Yang et al., 2025), and `GPT-4.1-mini` (Achiam et al., 2023). This selection allows us to compare models of different scales, architectural families, and training regimes, and to examine how such differences affect culturally grounded vision–language understanding across languages and dialects.

4.2 Prompting Techniques

To systematically evaluate the models' culturally grounded vision-language understanding, we em-

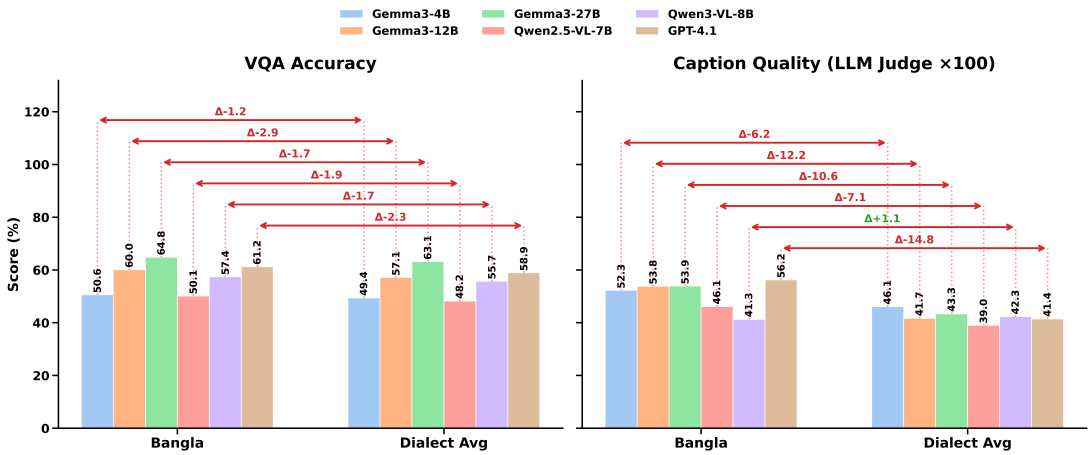

Figure 2: **Dialect variation across models.** We compare each model's average VQA accuracy (zero-shot, few-shot, and CoT) and caption quality on standard Bangla against its mean performance across the five regional dialects.

ploy zero-shot, few-shot, and chain-of-thought (CoT) (Wei et al., 2022) prompting strategies. The zero-shot and few-shot templates strictly instruct the models to generate outputs exclusively in the source language without extraneous explanations. For more complex reasoning in the VQA task, CoT prompting is utilized, allowing models to generate intermediate logical steps before yielding the final formatted answer. Across all evaluation settings, the decoding temperature is set to 0.1 to ensure highly deterministic outputs. All prompts are provided in Appendix C.

4.3 Evaluation Metrics

We evaluate the benchmark using task-specific metrics tailored to captioning and visual question answering. For **image captioning**, we report BERTScore-F1 (Zhang et al., 2020) to measure semantic similarity between generated and reference captions, and LLM-as-a-Judge (Gu et al., 2024) scores obtained with `Gemini-2.5-Flash` to capture overall caption quality beyond surface-level lexical overlap. Specifically, the judge is instructed to evaluate the captions across four dimensions: Relevance, Clarity, Conciseness, and Creativity, to compute a final holistic score ranging from 0 to 1. For **visual question answering**, we use accuracy (%), defined as the percentage of questions for which the model selects the correct answer option. To assess the reliability of the LLM-as-a-Judge evaluation, we conducted a focused human review with two independent subject-matter experts (SMEs) fluent in Bangla and experienced in culturally grounded language interpretation. Inter-annotator agreement between the two SMEs was measured over the evaluated samples using Cohen's κ, yielding a score of 0.78, which indicates strong agreement. We designed a 0–100 human scoring rubric with three components: (i) *decision accuracy* (0–50 points), which evaluates whether the judge makes the correct quality judgment; (ii) *reasoning alignment* (0–40 points), which measures whether the supporting analysis is logically consistent with the decision; and (iii) *explanation clarity* (0–10 points), which captures how clear and well-justified the explanation is. Using this rubric, the SMEs reviewed 100 output samples and ensured consistent scoring across cases. The LLM-as-a-Judge achieved a Human Consistency Score ranging from 90.89 to 92.01 out of 100 across experiments, indicating that its judgments align closely with expert evaluation.

5 Results and Analysis

In this section, we analyze the performance of multilingual VLMs through three core research questions (RQs). For the full experimental results please refer to Appendix E.

RQ1: How robust are multilingual VLMs to dialectal variation when understanding the same Bengali cultural content? To answer RQ1, we compare standard Bangla with the five Bangla dialects, as shown in Figure 2, while keeping the underlying image fixed so that any gap reflects linguistic variation rather than visual difficulty. Overall, multilingual VLMs are **not fully robust to dialectal variation**. The drop is modest but consistent for VQA, and much larger for caption generation. For instance, `GPT-4.1-mini` drops from

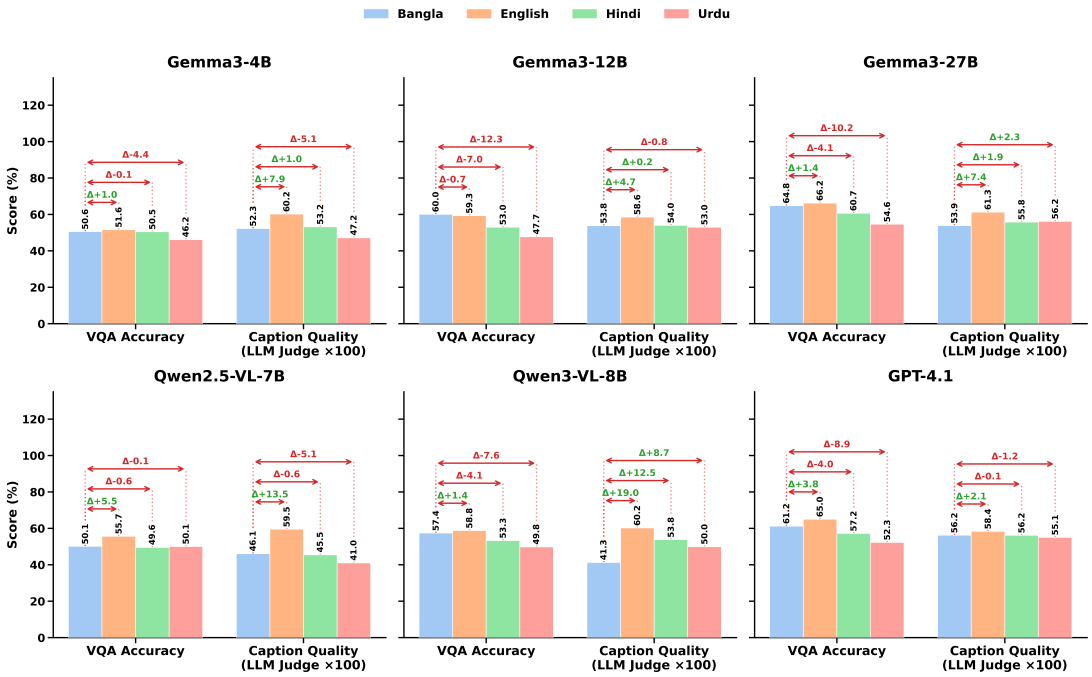

Figure 3: **Cross-lingual preservation of Bengali cultural meaning across models.** For each model, we compare each model's average VQA accuracy (zero-shot, few-shot, and CoT) and caption quality across 4 languages.

61.22% to 58.93% in VQA from Bangla to Dialect Avg (-2.29 points), but its caption quality falls from 56.24 to 41.41 on LLM-Judge$\times$100 (-14.84 points). A similar pattern holds for `Gemma3-27B`, which remains the strongest open model on this comparison: its VQA decreases only from 64.82% to 63.14% (-1.68 points), while caption quality drops from 53.90 to 43.35 (-10.55 points). These results show that dialectal robustness is substantially weaker for free-form generation than for answer-constrained reasoning.

The dialect penalty is also model-dependent, but no model is fully dialect-invariant. `Gemma3-12B` shows a -2.90-point VQA drop and a -12.15-point caption drop, while `Qwen2.5-VL-7B` drops by -1.91 VQA points and -7.10 caption points. Even when the VQA degradation is small, the captioning degradation remains substantial, indicating that models can often still identify the correct answer from options but struggle to generate culturally grounded descriptions when the input is phrased in dialect. One exception is `Qwen3-VL-8B`, whose caption score is nearly flat between Bangla and Dialect Avg (41.26 vs. 42.32), although its overall caption quality remains much lower than the strongest models. Taken together, these findings answer RQ1 clearly: current multilingual VLMs show only limited robustness to Bangla dialect variation, and evaluation on standard Bangla alone would overestimate their true cultural understanding.

RQ2. Do historically linked languages preserve Bengali cultural meaning better than standard cross-lingual translation would suggest? To answer RQ2, we compare the four standard-language settings, *Bangla*, *English*, *Hindi*, and *Urdu*, to test whether historically linked languages preserve Bengali cultural meaning better than a simple translation baseline would suggest. Figure 3 presents the comparative results across the four standard-language settings. The results show a mixed but informative pattern: **English remains the strongest overall transfer language**, but **Hindi and, to a lesser extent, Urdu often preserve cultural meaning better in caption generation than their lower VQA scores alone would suggest**. Averaged across models, English achieves the best performance (59.44% zero-shot VQA; 59.69 LLM-Judge$\times$100), followed by Bangla (57.37%; 50.59), Hindi (54.04%; 53.08), and Urdu (50.12%; 50.40). Notably, Hindi's average caption quality (53.08) is higher than Bangla's (50.59), even though its VQA is lower, suggesting that historically linked languages can retain culturally meaningful descriptive content even when answer selection remains harder.

This pattern is especially visible in representa-

tive strong models. For `GPT-4.1-mini`, caption quality is almost identical between Bangla and Hindi (56.24 vs. 56.19), despite a larger VQA gap (61.22% vs. 57.19%). For `Gemma3-27B`, Urdu caption quality (56.21) even exceeds its Bangla score (53.90), while VQA still drops from 64.82% to 54.62%. These results indicate that historically linked languages do preserve part of Bengali cultural meaning, particularly in semantic and descriptive generation, but this advantage is not uniform across tasks. In short, the answer to RQ2 is **partially yes**: Hindi and Urdu retain more Bengali cultural meaning than a naive cross-lingual view might expect, especially for captioning, but they do not fully match Bangla or English in structured cultural reasoning.

RQ3. Is the main bottleneck in Bengali culture understanding visual grounding, linguistic variation, or missing cultural knowledge? To answer RQ3, we average results by domain across all models and all language/dialect variants, and compare three signals: *zero-shot VQA*, *CoT VQA*, and *caption quality*. To make the ranking explicit and reproducible, we define a heuristic **difficulty score** function, $f(d)$, where d represents a specific domain data point. This function measures how far the results are from perfect (100%) by summing the error rate of the best-performing VQA setting and the error rate of caption generation:

$$f(d) = (100 - \max(\text{ZS}_d, \text{FS}_d, \text{CoT}_d)) + (100 - \text{Cap.}_d)$$

Intuitively, this formulation measures the absolute gap from optimal performance. A domain is ranked as more difficult if it consistently exhibits lower accuracy despite few-shot or CoT assistance, alongside poor caption generation. Table 3 suggests that **domain-level difficulty is strongly associated with knowledge demands, which may be a more significant factor than pure visual recognition.** Visually concrete domains such as Food are the easiest (difficulty 75.15), while culturally knowledge-intensive domains such as Media & Movies and Politics are the hardest (120.61 and 106.95). This spread indicates models struggle when interpretation depends on background knowledge, named entities, or culturally specific context. We note $f(d)$ is a heuristic proxy; low performance in Politics could also reflect weak entity recognition or answer-option confusability rather than a clean "knowledge versus grounding" distinction. Still, compared to much smaller language/dialect gaps, domain-level knowledge demands appear as the primary bottleneck.

R	Domain	ZS	FS	CoT	Cap.	Diff.	Type
1	Food	76.56	69.03	77.31	47.54	75.15	○
2	Cult.	63.42	59.01	65.48	49.23	85.29	△
3	Nat. Achv.	55.35	50.86	62.33	49.78	87.89	▲
4	Hist.	61.07	56.74	63.59	48.35	88.06	△
5	Pers.	58.67	57.42	62.28	43.45	94.27	▲
6	Sports	51.91	47.44	57.21	48.26	94.53	▲
7	Nature	52.29	53.05	52.71	50.58	96.37	○
8	Pol.	45.16	44.37	48.11	44.95	106.95	▲
9	M&M	33.63	31.93	36.27	43.12	120.61	▲

Table 3: **Estimating domain-level bottlenecks in Bengali culture understanding.** Domains are ranked from easier to harder using the heuristic difficulty score. Based on how domains need to be answered, we labeled them into three buckets: ○ = visually easier, △ = mixed, ▲ = knowledge-intensive. R = Rank, ZS = Zero-shot VQA, FS = Few-shot VQA, CoT = CoT VQA, Cap. = Caption Quality, Diff. = Difficulty Score.

This conclusion is reinforced by the reasoning trends. CoT yields the largest gains in knowledge-heavy domains like National Achievements (+6.97) and Sports (+5.31) compared to visually easier ones like Food (+0.75), though overall performance remains far from perfect. Conversely, few-shot prompting generally degrades VQA accuracy across most domains, suggesting models struggle to effectively leverage in-context cultural examples. Caption quality follows a similar pattern, with Media & Movies being the weakest (43.12). Taken together, these results support the hypothesis that models **often lack the specific cultural knowledge and fine-grained entity recognition needed to interpret what they see**. While linguistic variation poses a secondary challenge, domain-level difficulty indicates that cultural knowledge demands represent the most substantial bottleneck.

6 Conclusion

In this paper, we introduce BanglaVerse, a multilingual and multidialectal benchmark evaluating VLM performance on Bengali culture. Through a curated dataset of ∼32.3K artifacts tested across four languages and five regional dialects, we demonstrate that current VLMs are highly sensitive to linguistic variation. The observed performance drops highlight that the core limitation of these models lies in insufficient cultural knowledge rather than pure visual grounding.

Limitations

BanglaVerse offers a culturally grounded benchmark that, by design, prioritizes depth of cultural representation across nine carefully curated domains. While the current dataset comprises 1,152 expert-annotated images, this scale was intentionally chosen to ensure high annotation quality and cultural authenticity. Future iterations can naturally expand both in volume and modality coverage as the benchmark evolves.

Our annotation pipeline involved rigorous cross-verification by native speakers and cultural experts, which helps minimize subjective bias. As with any culturally rich dataset, certain local nuances may benefit from further community-driven refinement over time. Reassuringly, our systematic human evaluation did not surface notable inconsistencies in dialectal translations, lending confidence to the current annotations.

Finally, our evaluation covers a representative set of recent multilingual vision-language models. Given the rapid pace of model development, we view our results as establishing a reliable baseline for current capabilities, and we anticipate that BanglaVerse will serve as a continuing testbed as new architectures emerge.

Ethical Considerations

The dataset was developed with careful attention to ethical standards. All images were collected from publicly available sources and contain no personally identifiable information. Annotations were manually cross-verified to minimize bias, ensure cultural sensitivity, and avoid harmful or offensive content. The dataset is released solely for research purposes to advance multimodal understanding in Bangla, and we encourage responsible use that respects cultural contexts and does not promote misuse or discrimination.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

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, and 25 others. 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.

Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Farhan Ishmam, Fabiha Haider, Fariha Tanjim Shifat, Md Fahim, and Md. Farhad Alam. 2024. Chitrojera: A regionally relevant visual question answering dataset for bangla. *CoRR*, abs/2410.14991.

Md Fahim, Md Sakib Ul Rahman, Akm Moshiur Rahman, Md Farhan Ishmam, Md Tasmim Rahman, Fariha Tanjim Shifat, Fabiha Haider, and Md Farhad Alam Bhuiyan. 2026. Banglaprotha: Evaluating vision language models in underrepresented long-tail cultural contexts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1159–1169.

Ali Faraz, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami, Abhinav Ravi, Chandra Khatri, Shubham Agarwal, and 1 others. 2025. Indicvisionbench: Benchmarking cultural and multilingual understanding in vlms. *arXiv preprint arXiv:2511.04727*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Mahmud Hasan, Labiba Islam, Jannatul Ruma, Tasmiah Mayeesha, and Rashedur Rahman. 2023. Visual question generation in Bengali. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 10–19, Prague, Czech Republic. Association for Computational Linguistics.

Sheikh Mehedi Hasan and Adilur Rahaman. 2014. Standard dialect ideology in bangladesh: A field study. *Language in India*, 14(10).

SM Sajid Hasan, Shifat Islam, Mahamudul Hasan Rafi, S.M. Hasan Imtiaz Labib, and Faisal Muhammad Shah. 2025. Bengalivqa: A benchmark dataset for

bengali visual question answering. https://doi.org/10.17632/y9fw6k37n9.1. Dataset.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4.

Protiva Rani Karmaker. 2019. Dialectical and linguistic variations of bangla sounds: Phonemic analysis. *Jagannath University Journal of Arts*, 9(2):125–130.

Amr Keleg. 2025. Llm alignment for the arabs: A homogenous culture or diverse ones? In *The 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, page 1.

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

Mehraj Hossain Mahi, Anzir Rahman Khan, and Mayen Uddin Mojumdar. 2025. Bangladial: A merged and imbalanced text dataset for bengali regional dialect analysis. *Data in Brief*, page 112200.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. *arXiv preprint arXiv:2509.19274*.

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.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint*.

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

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems (NeurIPS 2024) — Datasets and Benchmarks Track*. Curran Associates, Inc.

Abu Shahen, Bellal Hossain, Md Bokul Hossain, and Most Nushrat Jahan. 2019. Globalization and bangladesh: An analysis from cultural perspective. *IOSR Journal of Humanities and Social Science*, 25(1):32–41.

Willem Van Schendel. 2020. *A history of Bangladesh*. Cambridge University Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Details of the Fine-Tuning Model

To systematically generate multidialectal variants of the BanglaVerse benchmark, we fine-tuned the `Qwen2.5-3B-Instruct` large language model. The model was specifically adapted to translate standard Bangla textual artifacts into five distinct regional dialects (Barishal, Chittagong, Noakhali, Rangpur, and Sylhet) and to natively handle dialect-based conversational generation.

A.1 Dataset Preparation and Prompt Formatting

The fine-tuning dataset was derived from the BanglaDial corpus and structured into two primary instruction formats to teach the model both direct translation and context-aware generation.

Task 1: Standard-to-Dialect Translation. For caption and direct context translation, the model was fed pairs of standard Bengali and their dialectal equivalents using the following structured prompt:

```
Translate the following standard Bengali to [Dialect] dialect:
Standard Bengali: [Standard Text]
[Dialect] dialect: [Target Dialect Text]
```

Task 2: Native Dialect Conversation. To ensure the model could process and answer questions formatted in native dialects (essential for the VQA task), we formulated conversational QA pairs:

```
User: [Question in Dialect]
Assistant: [Answer in Dialect]
```

During tokenization, inputs and target outputs were concatenated, and the causal language modeling objective was applied across the sequence. Tokenized sequences were capped at a maximum length of 512 tokens.

A.2 Memory Optimization and PEFT Configuration

Given the substantial parameter count of the base model, we utilized Low-Rank Adaptation (LoRA) alongside several memory-saving techniques to stabilize training on standard GPU infrastructure.

Prior to applying LoRA, we disabled the KV-cache (`use_cache=False`) and explicitly enabled gradient checkpointing to significantly reduce the VRAM footprint. LoRA adapters were subsequently injected into all primary linear projection layers within the transformer blocks to maximize adaptation capability. The target modules included the attention matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the feed-forward network gates (`gate_proj`, `up_proj`, `down_proj`). The LoRA rank (r) was set to 16 with a scaling factor (α) of 32 and a dropout rate of 0.1.

A.3 Training Dynamics and Hyperparameters

Training was executed using mixed precision (FP16). To mitigate memory constraints while maintaining a stable gradient update, we utilized a micro-batch size of 1 per device, compensated by 8 gradient accumulation steps.

To further optimize training speed, we enabled sequence length grouping (`group_by_length=True`), which minimizes the amount of padding required per batch. The data collator was configured to pad sequences to a multiple of 8, optimizing tensor core utilization on the GPU. The comprehensive hyperparameter configuration is summarized in Table 4.

A.4 Convergence and Inference Generation

The model converged successfully over the course of 3 epochs, completing exactly 8,600 global training steps. Logging was captured every 50 steps, with checkpoints saved every 1,000 steps. At the final training step, the model achieved a training loss of **0.4528**, with a gradient norm of 1.8488 and a decayed learning rate of approximately 1.41×10^{-7}. For the generation of the final dialectal artifacts used in the benchmark, the fine-tuned model was deployed with a temperature of 0.7 and sampling enabled

Hyperparameter	Value
Base Architecture	`Qwen/Qwen2.5-3B-Instruct`
Task Type	Causal Language Modeling
Training Epochs	3
Max Sequence Length	512
Per-Device Batch Size	1
Gradient Accumulation Steps	8
Learning Rate	5×10^{-5}
Warmup Steps	500
Precision	FP16
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Target Modules	All linear projection layers
LoRA Dropout	0.1
Sequence Grouping	Enabled

Table 4: Comprehensive hyperparameters and configurations used for fine-tuning the dialect generation model.

(`do_sample=True`) to ensure fluent, naturally varying dialectal structures, generating up to 50 new tokens per artifact.

B Human Evaluation Details

To ensure the multilingual and multidialectal artifacts in BanglaVerse maintained high semantic fidelity and cultural authenticity, we conducted a targeted human evaluation. This appendix provides a detailed breakdown of the evaluation methodology, the score distribution, and a brief error analysis.

B.1 Evaluation Setup and Guidelines

We randomly sampled 50 instances (comprising both the generated caption and the VQA pair) from each of the 8 generated linguistic variants (3 translated languages and 5 Bangla dialects), resulting in a total of 400 evaluated instances. Standard Bangla was excluded from this validation step because it serves as the manually curated source text rather than a generated output.

The evaluation was conducted by two distinct groups. Three of the authors evaluated the English, Hindi, and Urdu generations using a back-translation and cross-referencing approach to verify semantic and cultural preservation. Concurrently, five undergraduate students, selected specifically because each is a native speaker of one of the five target dialects (Barishal, Chittagong, Noakhali, Rangpur, and Sylhet), evaluated the dialectal generations to ensure naturalness and regional authenticity. Evaluators were instructed to rate each instance using a strict 3-point scale, which is defined in Table 5.

Score	Rating Name	Description
0	Inaccurate / Unusable	The translation alters the core meaning of the original Bangla, hallucinates information, or completely fails to capture the target dialect's vocabulary and grammar.
1	Acceptable with Minor Revisions	The core meaning is preserved, but the phrasing is slightly unnatural, or a specific cultural entity was translated too literally, requiring minor manual correction.
2	Flawless / Highly Authentic	The translation is perfectly natural, grammatically correct, culturally accurate, and (for dialects) reads exactly as a native speaker would converse.

Table 5: The 3-point grading scale used by human annotators to evaluate the generated artifacts.

B.2 Score Distribution and Inter-Annotator Agreement

Each of the 400 sampled instances was independently reviewed. To measure consistency, 20% of the samples were scored by multiple annotators, resulting in a strong Cohen's Kappa (κ) of 0.89, indicating near-perfect agreement.

The vast majority of the generated artifacts received perfect scores, validating the effectiveness of our LLM translation pipeline and our fine-tuned `Qwen2.5-3B-Instruct` dialect model. The absolute breakdown of the 400 evaluated samples is presented in Table 6.

Score Rating	Count	Percentage
2 (Flawless / Highly Authentic)	376	94.0%
1 (Acceptable with Minor Revisions)	20	5.0%
0 (Inaccurate / Unusable)	4	1.0%
Total Evaluated Instances	**400**	**100%**

Table 6: Distribution of human evaluation scores across 400 randomly sampled artifacts (50 from each of the 8 generated linguistic variants).

C Prompting Strategies: Templates and Analysis

C.1 Prompt Templates

Zero-shot Prompt: Caption

```
You are an assistant who generates short, fluent captions in [SOURCE LANGUAGE] only.
Look carefully at the given image and write exactly one meaningful sentence describing
it.
Do not use any [OTHER LANGUAGE] words, do not add extra explanations, labels, or quotes.
Your entire output must be only the Bangla caption as plain text.
```

Few-shot Prompt: Caption

```
You are an assistant who generates short, fluent captions in [SOURCE LANGUAGE] only.
Here are three example captions:
Example 1: Image: {example_image_1} {example_caption_1}
Example 2: Image: {example_image_2} {example_caption_2}
Example 3: Image: {example_image_3} {example_caption_3}
Now, generate a caption for the following image: Image: {image_path}
Write exactly one meaningful [SOURCE LANGUAGE] sentence. Do not use any [OTHER LANGUAGE]
words, do not add extra explanations, labels, or quotes. Your entire output must be only
the [SOURCE LANGUAGE] caption as plain text.
```

Zero-shot Prompt: VQA

```
You are an AI assistant that answers visual multiple-choice questions in [SOURCE
LANGUAGE].
Task:
1. Look carefully at the given image: {image_path}
2. Read the question: {question}
3. Review the provided answer choices: {options}
4. Select the single most accurate answer.
Response Rules:
- The index must be the programming list index (starting from 0).
- Respond ONLY with the exact format below.
- Use [SOURCE LANGUAGE] text for the answer option.
- Do NOT add explanations, extra words, reasoning steps, or anything outside the specified
format.
- Follow this exact structure:
Index: <option_index>, Answer: "<option_text_in_[SOURCE LANGUAGE]>"
```

Few-shot Prompt: VQA

```
You are an AI assistant that answers visual multiple-choice questions in [SOURCE
LANGUAGE].
Task:
1. Look carefully at the given image: {image_path}
```

2. Read the question: {question}
3. Review the provided answer choices: {options}
4. Select the single most accurate answer.
Response Rules:
- The index must be the programming list index (starting from 0).
- Respond ONLY with the exact format below.
- Use [SOURCE LANGUAGE] text for the answer option.
- Do NOT add explanations, extra words, reasoning steps, or anything outside the specified format.
- Follow this exact structure: Index: <option_index>, Answer: "<option_text_in_[SOURCE LANGUAGE]>"
Here are three example QA pairs:
Example 1: Image: {example_image_1} Question: {example_question_1} Options: {example_options_1} Answer: Index: {example_index_1}, Answer: "{example_answer_1}"
Example 2: Image: {example_image_2} Question: {example_question_2} Options: {example_options_2} Answer: Index: {example_index_2}, Answer: "{example_answer_2}"
Example 3: Image: {example_image_3} Question: {example_question_3} Options: {example_options_3} Answer: Index: {example_index_3}, Answer: "{example_answer_3}"
Now, answer for the given image. Image: {image_path} Question: {question} Options: {options} Answer:

CoT Prompt: VQA

You are an AI assistant that answers visual multiple-choice questions in [SOURCE LANGUAGE].
Task:
1. Look carefully at the given image: {image_path}
2. Read the question: {question}
3. Review the provided answer choices: {options}
4. Select the single most accurate answer.
Response Rules:
- The index must be the programming list index (starting from 0).
- Use [SOURCE LANGUAGE] text for the answer option.
- In Reasoning_En, write step-by-step reasoning in English - break down the solution logically:
Step 1: Describe key visual observations.
Step 2: Match observations to relevant answer choices.
Step 3: Eliminate incorrect choices with brief justification.
Step 4: Conclude why the final choice is correct.
- Be clear, concise, and factual (avoid overly long explanations).
- Follow this exact response format:
Reasoning_En:
Step 1: <your_observations>
Step 2: <your_matching_logic>
Step 3: <your_elimination_of _wrong_options>
Step 4: <your_final_choice _reasoning>
Final Answer: Index: <option_index>, Answer: "<option_text_in_[SOURCE LANGUAGE]>"

LLM-as-a-Judge Prompt

You are a highly skilled and impartial caption evaluator. Your task is to carefully compare a generated caption with a reference caption and score it according to the following dimensions:
1. Relevance (0-1): How well does the generated caption describe the main objects, actions, and context of the reference caption? Reward high semantic overlap and penalize missing or hallucinated details.
2. Clarity (0-1): Is the caption grammatically correct, well-structured, and easy to read?
3. Conciseness (0-1): Is the caption free of redundancy, filler words, or unnecessary complexity while still conveying the full meaning?
4. Creativity (0-1): Does the caption show originality or an engaging phrasing, rather than being overly generic?
After scoring each dimension, compute an Overall (0-1) score that reflects the holistic quality of the generated caption, giving slightly higher weight to Relevance and Clarity.
Your response must strictly follow this JSON-like structure:
Relevance: [float between 0 and 1]

```
Clarity: [float between 0 and 1]
Conciseness: [float between 0 and 1]
Creativity: [float between 0 and 1]
Overall: [float between 0 and 1]
Explanation: [Concise explanation: mention key strengths, weaknesses, and reasoning for
the scores.]
Reference Caption: {reference_caption}
Generated Caption: {generated_caption}
```

C.2 Additional Analysis

RQ4. How do different prompting strategies impact culturally grounded multimodal tasks?

Impact of Prompting Strategies Across Models. We investigate how large vision-language models respond to three distinct prompting techniques: zero-shot, few-shot (in-context learning), and Chain-of-Thought (CoT). To isolate the effect of these strategies, we aggregate performance across all six evaluated models, nine cultural domains, and all nine linguistic variants (four standard languages and five Bangla dialects). The results reveal a striking divergence between reasoning-based prompting and example-based prompting in cultural contexts.

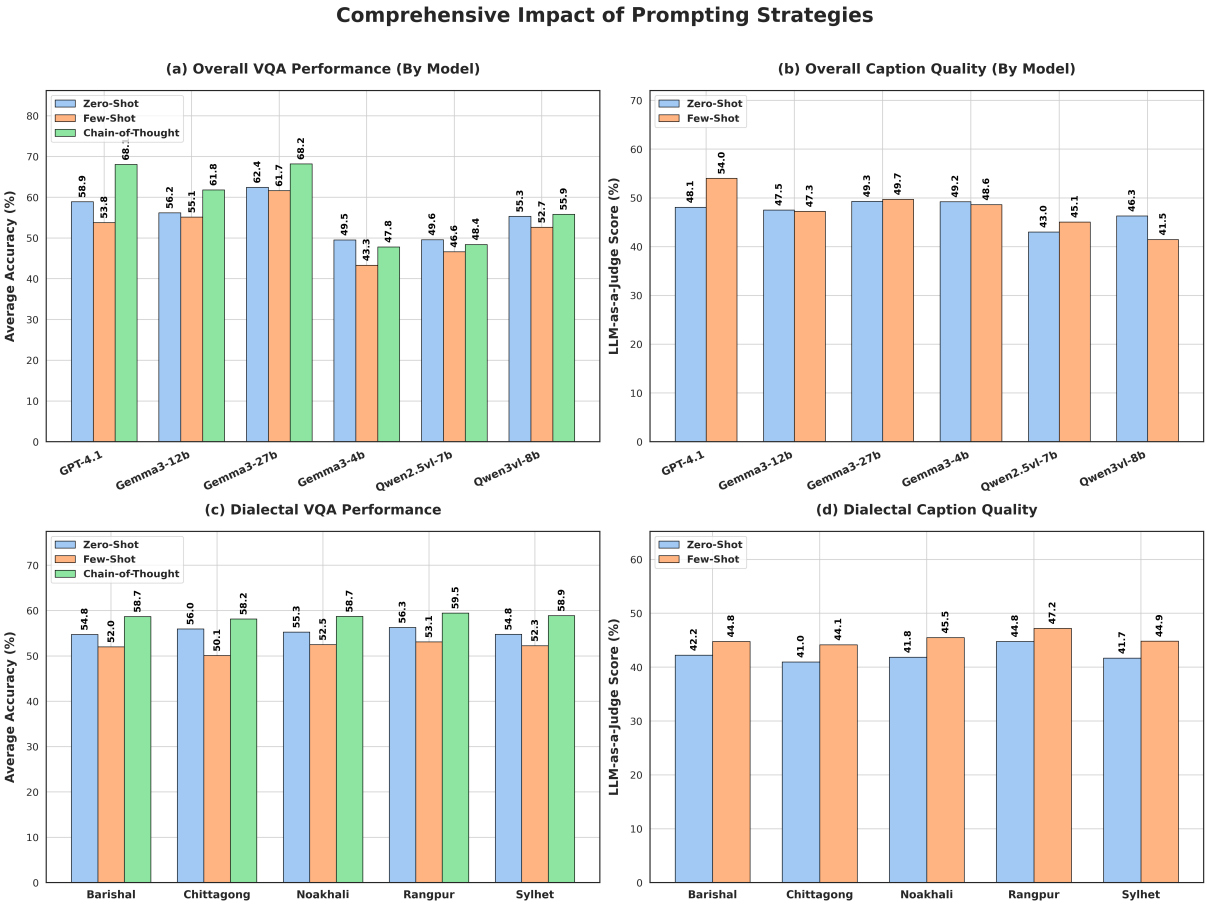

Figure 4: **Impact of Prompting Strategies on Culturally Grounded Tasks.** A unified comparison of Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) prompting. **(Top Row):** (a) Overall performance across all 9 linguistic variants and domains by evaluated models. CoT consistently enhances VQA (b) while Few-Shot degrades it, though Few-Shot slightly improves stylistic caption generation. **(Bottom Row):** (c) Performance specifically isolated across the five regional Bangla dialects. Dialects universally mirror the overall trend, with CoT enhancing discriminative tasks (d) and Few-Shot aiding in generating dialectal captions.

First, **explicit reasoning significantly enhances cultural understanding.** Compared to standard zero-shot VQA (55.34% average accuracy), applying CoT yields a consistent overall improvement, raising

average accuracy to 58.36%. This gain is most pronounced in highly capable models, as illustrated in Figure 4a. For instance, GPT-4.1 surges from 58.93% to 68.09% under CoT, an absolute gain of over 9%, while Gemma3-27B improves from 62.45% to 68.21%. By generating intermediate reasoning steps, models can successfully unpack visual evidence, recognize regional entities, and connect them to specific background knowledge before committing to an answer. This is particularly beneficial in knowledge-heavy domains such as *National Achievements*, which sees a jump from 55.35% to 62.33%.

Conversely, we observe a surprising phenomenon regarding in-context learning: **few-shot prompting consistently degrades VQA performance.** Across all evaluated models, few-shot VQA accuracy (52.20%) is markedly lower than zero-shot accuracy. For example, Gemma3-4B experiences a substantial 6.20% absolute drop, and even state-of-the-art models like GPT-4.1 suffer a 5.12% decline. In culturally knowledge-intensive tasks, few-shot examples appear to act as distractors rather than guides. We hypothesize that providing interleaved image-text examples from different cultural domains clutters the limited multimodal context window, inducing entity bias and causing models to hallucinate visual features from the demonstrations into the target image.

However, for generative tasks, the trend is more nuanced. While few-shot prompting harms discriminative VQA, it provides marginal benefits for open-ended captioning (Figure 4b), raising the overall LLM-as-a-Judge score from 47.25% to 47.69%. Notably, it yields significant gains in specific domains like *Media & Movies* (improving from 43.12% to 50.52%). This suggests that while in-context examples do not successfully inject missing factual knowledge, they effectively teach models the expected stylistic formatting, tone, and length required for culturally descriptive captions.

Impact of Prompting Strategies Across Regional Dialects. To understand if non-standard linguistic variations react differently to these frameworks, we isolate the performance of the five Bangla dialects (Barishal, Chittagong, Noakhali, Rangpur, and Sylhet). Figure 4 (c and d) demonstrates that the overall prompting trends hold consistently across every regional dialect, albeit with varying degrees of sensitivity.

For the VQA task, **CoT acts as a universal enhancer** for dialectal understanding. Rangpur proved to be the most resilient dialect overall, achieving the highest average accuracy under CoT (59.47%) compared to its zero-shot baseline (56.29%). Other dialects, such as Sylhet and Noakhali, experienced similar absolute gains of roughly 3.5% to 4.1% when reasoning steps were explicitly generated. This indicates that guiding models to ”think” before answering helps them better parse phonological and lexical variations in regional text.

Conversely, **few-shot prompting universally degraded VQA performance** for every single dialect. While the average zero-shot accuracy across all dialects stood at 55.41%, providing few-shot examples dragged the average down to 52.00%. Notably, *Chittagong* exhibited the most severe vulnerability to in-context distractors; its VQA accuracy plummeted from a strong 55.95% in zero-shot to just 50.11% in few-shot, a steep absolute drop of nearly 6%. This suggests that models struggle particularly hard to map cross-image visual entities when the text is written in highly divergent regional scripts.

For the generative captioning task, however, the dialects benefited universally from few-shot prompting. While the average zero-shot LLM-as-a-Judge score for dialects was low (42.29%), providing in-context examples consistently raised the average to 45.29%. *Barishal* and *Noakhali* saw notable improvements of roughly 2.5% to 3.6% in caption quality, reinforcing the hypothesis that few-shot examples successfully teach models the expected generative structure and tone, even when the target output must be written in a specific regional dialect.

D Failed Examples

In this section, we present a selection of failed examples that highlight common error modes encountered by the evaluated models. These failures illustrate various limitations in current multilingual VLMs, ranging from cultural misidentification and surface-level descriptive fallbacks to language constraint violations and implicit biases. Analyzing these specific instances provides deeper insight into the gaps between standard visual recognition and genuine, culturally grounded understanding.

Culture_009

Ground Truth: "বিশ্ব ইজতেমায় বিশ্বের নানা প্রান্ত থেকে আসা মুসল্লিদের ভিড় জমেছে তুরাগতীরে। *(en: Devotees from around the world have gathered at the bank of the Turag for Bishwa Ijtema.)*

`Gemma3-4B` **Generated Caption:** كثيرة“ خيام ظل في تجمعون الناس من كبيرة.” *(bn:* অনেক মানুষ বড় বড় তাঁবুর নিচে জড়ো হয়।*; en: Many people gather under large tents.)*

Explanation: The generated caption is factually correct but generic and switches to Arabic instead of Bangla. Although models were strictly prompted to generate in Bangla, they occasionally shifted to other culturally related languages (e.g., Arabic for Islamic gatherings) while still producing meaningful outputs.

Personality_026

Ground Truth: "হাসান আজিজুল হক — লেখক, ছোটগল্প ও উপন্যাসের মাধ্যমে মানুষের জটিল মানসিকতা ফুটিয়ে তুলেছেন।" *(en: Hasan Azizul Haque — Writer who has portrayed the complex human psyche through short stories and novels.)*

`GPT-4.1-mini` **Generated Caption:** "চশমা ও সাদা পোশাক পরিহিত একজন প্রবীণ ব্যক্তি বইয়ের তাকের সামনে চিন্তাশীল ভঙ্গিতে তাকিয়ে আছেন।" *(en: An elderly man wearing glasses and white clothes looks thoughtfully in front of a bookshelf.)*

Explanation: While the generated caption is factually correct in describing visual attributes, it fails to capture the person's identity and cultural significance. This shows a common model limitation: defaulting to surface-level descriptions when unable to recognize notable personalities.

Food_010

Ground Truth: "ফুচকা, ঝাল ও টক-মিষ্টি স্বাদের এক অসাধারণ সমন্বয়।" *(en: A wonderful combination of fuchka, salty, and sweet-sour flavours.)*

`Gemma3-27B` **Generated Caption:** "পানির মধ্যে ভেজানো কিছু মশলা, সবুজ পাতা, এবং নারকেল ছাড়াও একটি সুস্বাদু পানীয় এবং ভাজা খাবার সমন্বিত একটি সুস্বাদু ভারতীয় রাস্তার খাবার।" *(en: A delicious Indian street food consisting of some spices, green leaves, and coconut soaked in water, along with a delicious drink and fried food.)*

Explanation: The generated caption is overly descriptive and shifts the focus toward generic Indian street food rather than identifying the dish as "fuchka." While factually plausible, it misrepresents the key cultural context.

Politics_006

Question: "ছবির নেত্রী কোন দলের প্রধান?" *(en: Which party is the leader of the picture?)*
Ground Truth Answer: "বিএনপি" *(en: BNP)*
`Gemma3-4B` **Predicted Answer:** "শেখ হাসিনা" *(en: Sheikh Hasina)*

Explanation: The correct answer is BNP, but the model generated Sheikh Hasina, reflecting a strong political bias toward the Awami League. This not only demonstrates over-association with ruling party figures but also shows factual inconsistency, as Sheikh Hasina is a person and not the name of a political party.

Food_008

Ground Truth Caption: "সুস্বাদু শিঙারা, তেলমশলার সোনালি রঙে ভাজা।" *(en: Delicious shingara, fried to a golden color in oil and spices.)*
`Qwen2.5-VL-7B` **Generated Caption:** "সামুদ্রিক খাবার দৃশ্য।" *(en: Seafood scene.)*

Explanation: The model failed to recognize the Bangladeshi snack *shingara* and instead hallucinated a completely unrelated "seafood scene." This illustrates cultural and recognition gaps, where local foods are often misinterpreted as generic or foreign items.

Personality_004

Ground Truth Caption: "উস্তাদ শাহ আবদুল করিম হচ্ছেন বাংলাদেশের কিংবদন্তি সঙ্গীতশিল্পী, সুরকার, গীতিকার ও সঙ্গীত শিক্ষক।" *(en: Ustad Shah Abdul Karim is a legendary musician, composer, lyricist, and music teacher of Bangladesh.)*
`Qwen2.5-VL-7B` **Zero-shot Caption:** "আমার চিত্রে একজন প্রাচীন মহিলা দেখা যাচ্ছেন।" *(en: An ancient woman is seen in my picture.)*
`Qwen2.5-VL-7B` **Few-shot Caption:** "বৃদ্ধ মহিলার প্রাণের স্বাদ ও সাহস।" *(en: The old woman's zest for life and courage.)*

Explanation: The model repeatedly misclassified Shah Abdul Karim, a legendary male musician of Bangladesh, as a woman, reflecting a gender bias where long hair is incorrectly equated with female identity. This highlights both cultural unfamiliarity and gender stereotyping in caption generation.

Sports_010

Question: "ছবির খেলোয়াড় হামজা চৌধুরি কোন পজিশনে খেলেন?" (en: What position does the player in the picture, Hamza Chowdhury, play?)
Ground Truth: "মিডফিল্ডার" *(en: Midfielder)*
`Gemma3-4B` **Predicted:** ”batsman”

Explanation: The model violated the language constraint by answering in English and produced a factually incorrect label. “Batsman” is a cricket role, whereas the correct answer is the football position midfielder, indicating cross-sport confusion and weak factual grounding.

E Full Experimental Results

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.69	0.53	0.63	0.69	0.52	0.59	0.52
	Food	0.70	0.52	0.74	0.70	0.53	0.63	0.78
	Hist.	0.69	0.54	0.47	0.68	0.47	0.37	0.44
	M&M	0.66	0.50	0.30	0.84	0.52	0.32	0.38
	Nat. Achv.	0.70	0.56	0.31	0.67	0.50	0.29	0.36
	Nature	0.70	0.58	0.66	0.69	0.54	0.74	0.59
	Pers.	0.65	0.45	0.61	0.69	0.49	0.62	0.49
	Pol.	0.67	0.50	0.39	0.69	0.48	0.40	0.39
	Sports	0.68	0.52	0.44	0.68	0.51	0.26	0.48
Gemma3-12B	Cult.	0.70	0.55	0.68	0.69	0.52	0.70	0.72
	Food	0.71	0.53	0.83	0.71	0.52	0.82	0.89
	Hist.	0.69	0.54	0.62	0.68	0.48	0.63	0.75
	M&M	0.66	0.51	0.40	0.78	0.52	0.32	0.33
	Nat. Achv.	0.69	0.56	0.49	0.66	0.49	0.50	0.59
	Nature	0.70	0.58	0.63	0.69	0.56	0.59	0.66
	Pers.	0.64	0.49	0.67	0.68	0.48	0.70	0.74
	Pol.	0.66	0.51	0.55	0.68	0.48	0.56	0.61
	Sports	0.68	0.57	0.53	0.67	0.48	0.47	0.46
Gemma3-27B	Cult.	0.67	0.44	0.72	0.70	0.53	0.73	0.73
	Food	0.71	0.58	0.87	0.72	0.55	0.89	0.89
	Hist.	0.70	0.56	0.69	0.69	0.51	0.69	0.74
	M&M	0.67	0.47	0.36	0.82	0.53	0.32	0.42
	Nat. Achv.	0.71	0.61	0.51	0.67	0.50	0.51	0.78
	Nature	0.70	0.57	0.66	0.70	0.59	0.72	0.78
	Pers.	0.65	0.51	0.78	0.70	0.52	0.83	0.76
	Pol.	0.67	0.52	0.56	0.69	0.49	0.50	0.62
	Sports	0.70	0.58	0.69	0.70	0.52	0.60	0.74
Qwen2.5-VL-7B	Cult.	0.69	0.46	0.55	0.66	0.48	0.47	0.59
	Food	0.70	0.42	0.68	0.64	0.55	0.54	0.68
	Hist.	0.70	0.51	0.60	0.67	0.57	0.52	0.52
	M&M	0.70	0.44	0.22	0.76	0.50	0.27	0.17
	Nat. Achv.	0.69	0.46	0.69	0.69	0.51	0.71	0.54
	Nature	0.70	0.49	0.42	0.55	0.38	0.64	0.39
	Pers.	0.65	0.44	0.47	0.69	0.48	0.46	0.56
	Pol.	0.66	0.46	0.39	0.69	0.48	0.32	0.41
	Sports	0.68	0.47	0.48	0.72	0.52	0.48	0.40
Qwen3-VL-8B	Cult.	0.67	0.44	0.68	0.67	0.42	0.68	0.62
	Food	0.68	0.42	0.70	0.69	0.47	0.66	0.66
	Hist.	0.67	0.44	0.69	0.67	0.39	0.69	0.78
	M&M	0.67	0.36	0.40	0.67	0.37	0.28	0.38
	Nat. Achv.	0.66	0.40	0.78	0.65	0.37	0.63	0.70
	Nature	0.68	0.43	0.46	0.64	0.54	0.47	0.52
	Pers.	0.63	0.41	0.51	0.65	0.45	0.52	0.57
	Pol.	0.64	0.40	0.45	0.68	0.43	0.48	0.45
	Sports	0.66	0.41	0.49	0.64	0.34	0.60	0.45
GPT-4.1-mini	Cult.	0.71	0.57	0.65	0.71	0.57	0.56	0.77
	Food	0.72	0.56	0.88	0.73	0.54	0.81	0.94
	Hist.	0.71	0.58	0.66	0.72	0.57	0.60	0.72
	M&M	0.70	0.54	0.49	0.89	0.61	0.42	0.46
	Nat. Achv.	0.71	0.62	0.58	0.71	0.60	0.57	0.78
	Nature	0.71	0.60	0.57	0.72	0.61	0.56	0.65
	Pers.	0.66	0.51	0.65	0.71	0.53	0.60	0.75
	Pol.	0.66	0.50	0.50	0.69	0.52	0.50	0.61
	Sports	0.69	0.58	0.53	0.71	0.55	0.48	0.69

Table 7: Full results by model and domain for Bangla Language.

Models	Domain	Zero-Shot Caption		Zero-Shot VQA	Few-Shot Caption		Few-Shot VQA	CoT VQA
		B-F1	LLM	Accuracy (%)	B-F1	LLM	Accuracy (%)	Accuracy (%)
Gemma3-4B	Cult.	0.87	0.65	0.72	0.86	0.56	0.58	0.50
	Food	0.85	0.61	0.75	0.86	0.54	0.70	0.68
	Hist.	0.86	0.61	0.43	0.84	0.44	0.40	0.44
	M&M	0.85	0.55	0.26	0.92	0.55	0.34	0.42
	Nat. Achv.	0.87	0.63	0.41	0.87	0.54	0.34	0.33
	Nature	0.86	0.66	0.55	0.86	0.59	0.55	0.39
	Pers.	0.83	0.55	0.57	0.84	0.51	0.52	0.56
	Pol.	0.84	0.56	0.41	0.83	0.49	0.43	0.44
	Sports	0.86	0.60	0.54	0.84	0.43	0.42	0.55
Gemma3-12B	Cult.	0.87	0.62	0.69	0.86	0.55	0.72	0.71
	Food	0.85	0.59	0.73	0.86	0.54	0.77	0.73
	Hist.	0.85	0.60	0.62	0.85	0.51	0.60	0.74
	M&M	0.85	0.53	0.44	0.90	0.59	0.30	0.40
	Nat. Achv.	0.87	0.63	0.57	0.86	0.48	0.50	0.60
	Nature	0.86	0.64	0.53	0.85	0.56	0.58	0.64
	Pers.	0.83	0.53	0.62	0.85	0.50	0.71	0.65
	Pol.	0.83	0.54	0.52	0.85	0.50	0.52	0.63
	Sports	0.85	0.58	0.61	0.85	0.50	0.56	0.62
Gemma3-27B	Cult.	0.87	0.66	0.77	0.86	0.57	0.75	0.80
	Food	0.85	0.62	0.88	0.86	0.54	0.84	0.82
	Hist.	0.85	0.65	0.62	0.87	0.57	0.62	0.73
	M&M	0.86	0.53	0.41	0.92	0.54	0.36	0.45
	Nat. Achv.	0.87	0.62	0.65	0.86	0.55	0.50	0.86
	Nature	0.86	0.67	0.60	0.85	0.55	0.61	0.68
	Pers.	0.83	0.55	0.81	0.85	0.50	0.79	0.77
	Pol.	0.83	0.55	0.51	0.84	0.51	0.47	0.49
	Sports	0.87	0.66	0.70	0.85	0.50	0.70	0.61
Qwen2.5-VL-7B	Cult.	0.87	0.69	0.62	0.84	0.59	0.65	0.58
	Food	0.85	0.60	0.77	0.81	0.62	0.72	0.75
	Hist.	0.86	0.58	0.62	0.83	0.52	0.54	0.60
	M&M	0.86	0.53	0.32	0.90	0.54	0.30	0.37
	Nat. Achv.	0.88	0.64	0.60	0.87	0.62	0.51	0.61
	Nature	0.87	0.68	0.49	0.77	0.56	0.40	0.48
	Pers.	0.83	0.54	0.51	0.84	0.51	0.52	0.60
	Pol.	0.83	0.55	0.49	0.84	0.54	0.42	0.56
	Sports	0.85	0.56	0.59	0.86	0.60	0.49	0.59
Qwen3-VL-8B	Cult.	0.87	0.63	0.71	0.86	0.50	0.66	0.64
	Food	0.85	0.62	0.78	0.83	0.58	0.73	0.71
	Hist.	0.83	0.62	0.64	0.86	0.49	0.61	0.63
	M&M	0.87	0.56	0.37	0.91	0.48	0.31	0.45
	Nat. Achv.	0.87	0.73	0.65	0.88	0.60	0.56	0.69
	Nature	0.85	0.61	0.46	0.81	0.43	0.38	0.47
	Pers.	0.83	0.54	0.64	0.85	0.49	0.64	0.63
	Pol.	0.82	0.52	0.45	0.85	0.48	0.45	0.42
	Sports	0.85	0.59	0.60	0.86	0.58	0.55	0.65
GPT-4.1-mini	Cult.	0.87	0.64	0.72	0.87	0.66	0.68	0.76
	Food	0.86	0.64	0.87	0.87	0.67	0.78	0.92
	Hist.	0.84	0.61	0.64	0.87	0.54	0.66	0.70
	M&M	0.85	0.43	0.51	0.95	0.66	0.36	0.51
	Nat. Achv.	0.88	0.66	0.68	0.88	0.69	0.55	0.77
	Nature	0.86	0.66	0.65	0.86	0.63	0.58	0.70
	Pers.	0.83	0.53	0.83	0.86	0.58	0.76	0.82
	Pol.	0.83	0.54	0.40	0.85	0.50	0.43	0.58
	Sports	0.86	0.55	0.56	0.87	0.61	0.46	0.75

Table 8: Full results by model and domain for English Language.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.71	0.57	0.72	0.69	0.52	0.53	0.58
	Food	0.71	0.53	0.70	0.71	0.52	0.57	0.65
	Hist.	0.70	0.55	0.52	0.69	0.47	0.32	0.52
	M&M	0.67	0.49	0.26	0.84	0.52	0.25	0.32
	Nat. Achv.	0.70	0.56	0.29	0.69	0.53	0.30	0.38
	Nature	0.70	0.58	0.66	0.71	0.56	0.61	0.34
	Pers.	0.65	0.48	0.57	0.68	0.47	0.45	0.45
	Pol.	0.67	0.50	0.36	0.69	0.45	0.41	0.40
	Sports	0.68	0.54	0.47	0.69	0.46	0.37	0.44
Gemma3-12B	Cult.	0.71	0.57	0.61	0.70	0.51	0.62	0.68
	Food	0.70	0.53	0.73	0.71	0.52	0.67	0.77
	Hist.	0.70	0.54	0.62	0.69	0.48	0.59	0.71
	M&M	0.70	0.53	0.40	0.81	0.51	0.34	0.35
	Nat. Achv.	0.71	0.58	0.52	0.69	0.50	0.37	0.68
	Nature	0.70	0.57	0.46	0.71	0.56	0.57	0.56
	Pers.	0.65	0.49	0.54	0.69	0.50	0.57	0.62
	Pol.	0.67	0.51	0.43	0.69	0.45	0.49	0.47
	Sports	0.68	0.54	0.47	0.69	0.49	0.45	0.52
Gemma3-27B	Cult.	0.72	0.58	0.71	0.71	0.53	0.73	0.73
	Food	0.71	0.56	0.84	0.71	0.54	0.81	0.87
	Hist.	0.70	0.54	0.67	0.70	0.50	0.63	0.72
	M&M	0.69	0.52	0.38	0.86	0.53	0.33	0.42
	Nat. Achv.	0.73	0.61	0.54	0.70	0.53	0.55	0.82
	Nature	0.71	0.59	0.61	0.70	0.57	0.57	0.62
	Pers.	0.66	0.50	0.66	0.70	0.50	0.68	0.68
	Pol.	0.68	0.52	0.49	0.69	0.49	0.49	0.45
	Sports	0.71	0.58	0.56	0.69	0.51	0.53	0.53
Qwen2.5-VL-7B	Cult.	0.70	0.47	0.57	0.64	0.43	0.56	0.60
	Food	0.71	0.41	0.64	0.65	0.36	0.67	0.63
	Hist.	0.70	0.50	0.58	0.68	0.38	0.48	0.55
	M&M	0.70	0.39	0.29	0.80	0.43	0.22	0.36
	Nat. Achv.	0.70	0.48	0.75	0.69	0.47	0.58	0.47
	Nature	0.71	0.49	0.41	0.65	0.48	0.50	0.45
	Pers.	0.65	0.43	0.33	0.69	0.48	0.43	0.45
	Pol.	0.67	0.46	0.40	0.69	0.47	0.34	0.41
	Sports	0.69	0.48	0.48	0.72	0.54	0.42	0.46
Qwen3-VL-8B	Cult.	0.71	0.58	0.60	0.68	0.40	0.52	0.64
	Food	0.71	0.53	0.72	0.67	0.34	0.65	0.75
	Hist.	0.71	0.50	0.66	0.68	0.43	0.71	0.70
	M&M	0.67	0.49	0.37	0.78	0.41	0.29	0.37
	Nat. Achv.	0.71	0.61	0.66	0.69	0.47	0.70	0.76
	Nature	0.70	0.60	0.36	0.60	0.38	0.40	0.47
	Pers.	0.65	0.49	0.48	0.69	0.50	0.48	0.49
	Pol.	0.66	0.51	0.43	0.69	0.47	0.45	0.43
	Sports	0.69	0.55	0.53	0.72	0.50	0.50	0.47
GPT-4.1-mini	Cult.	0.72	0.59	0.64	0.72	0.61	0.61	0.71
	Food	0.72	0.57	0.85	0.73	0.59	0.70	0.83
	Hist.	0.71	0.55	0.70	0.72	0.55	0.59	0.68
	M&M	0.69	0.52	0.38	0.87	0.58	0.37	0.42
	Nat. Achv.	0.72	0.61	0.59	0.72	0.56	0.64	0.73
	Nature	0.71	0.62	0.49	0.71	0.59	0.38	0.53
	Pers.	0.65	0.50	0.60	0.71	0.53	0.50	0.72
	Pol.	0.68	0.53	0.46	0.69	0.52	0.53	0.49
	Sports	0.69	0.56	0.44	0.74	0.54	0.44	0.53

Table 9: Full results by model and domain for Hindi Language.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.70	0.48	0.57	0.69	0.48	0.34	0.48
	Food	0.70	0.46	0.63	0.70	0.47	0.52	0.56
	Hist.	0.69	0.49	0.51	0.70	0.44	0.37	0.45
	M&M	0.69	0.45	0.19	0.85	0.52	0.24	0.28
	Nat. Achv.	0.69	0.49	0.31	0.69	0.46	0.26	0.38
	Nature	0.70	0.49	0.67	0.69	0.51	0.57	0.41
	Pers.	0.65	0.45	0.52	0.68	0.48	0.43	0.47
	Pol.	0.67	0.47	0.37	0.70	0.47	0.37	0.41
	Sports	0.68	0.47	0.39	0.70	0.48	0.34	0.52
Gemma3-12B	Cult.	0.71	0.54	0.56	0.69	0.50	0.64	0.60
	Food	0.69	0.51	0.67	0.70	0.51	0.65	0.68
	Hist.	0.70	0.53	0.59	0.70	0.48	0.59	0.63
	M&M	0.72	0.52	0.26	0.83	0.53	0.28	0.32
	Nat. Achv.	0.70	0.55	0.55	0.70	0.50	0.43	0.64
	Nature	0.70	0.56	0.36	0.70	0.56	0.47	0.46
	Pers.	0.66	0.49	0.47	0.69	0.50	0.61	0.55
	Pol.	0.68	0.53	0.41	0.70	0.47	0.50	0.50
	Sports	0.68	0.54	0.42	0.70	0.50	0.52	0.50
Gemma3-27B	Cult.	0.71	0.58	0.61	0.70	0.52	0.68	0.68
	Food	0.71	0.56	0.72	0.71	0.53	0.77	0.78
	Hist.	0.71	0.58	0.65	0.70	0.49	0.64	0.70
	M&M	0.73	0.51	0.32	0.87	0.52	0.32	0.35
	Nat. Achv.	0.71	0.61	0.58	0.70	0.54	0.57	0.83
	Nature	0.70	0.59	0.41	0.70	0.56	0.57	0.51
	Pers.	0.66	0.49	0.56	0.69	0.51	0.61	0.59
	Pol.	0.68	0.54	0.45	0.71	0.49	0.46	0.48
	Sports	0.71	0.60	0.61	0.73	0.49	0.60	0.61
Qwen2.5-VL-7B	Cult.	0.69	0.43	0.58	0.65	0.42	0.54	0.55
	Food	0.69	0.34	0.66	0.64	0.32	0.56	0.60
	Hist.	0.70	0.46	0.58	0.69	0.41	0.54	0.54
	M&M	0.66	0.35	0.24	0.80	0.41	0.29	0.35
	Nat. Achv.	0.68	0.40	0.67	0.71	0.46	0.65	0.55
	Nature	0.69	0.43	0.49	0.56	0.32	0.24	0.46
	Pers.	0.65	0.38	0.43	0.69	0.45	0.51	0.44
	Pol.	0.68	0.45	0.42	0.70	0.47	0.35	0.36
	Sports	0.67	0.46	0.44	0.72	0.44	0.29	0.50
Qwen3-VL-8B	Cult.	0.70	0.53	0.53	0.67	0.40	0.59	0.57
	Food	0.69	0.49	0.66	0.66	0.37	0.60	0.64
	Hist.	0.71	0.54	0.66	0.67	0.38	0.60	0.68
	M&M	0.68	0.37	0.25	0.82	0.47	0.39	0.26
	Nat. Achv.	0.71	0.57	0.69	0.69	0.50	0.50	0.70
	Nature	0.69	0.52	0.37	0.59	0.35	0.46	0.38
	Pers.	0.64	0.43	0.42	0.69	0.51	0.45	0.44
	Pol.	0.67	0.47	0.39	0.69	0.47	0.41	0.44
	Sports	0.69	0.58	0.51	0.72	0.47	0.46	0.49
GPT-4.1-mini	Cult.	0.72	0.57	0.64	0.72	0.57	0.60	0.69
	Food	0.71	0.54	0.69	0.72	0.56	0.61	0.78
	Hist.	0.72	0.55	0.60	0.71	0.56	0.56	0.63
	M&M	0.70	0.50	0.43	0.87	0.55	0.35	0.43
	Nat. Achv.	0.71	0.60	0.58	0.71	0.58	0.56	0.73
	Nature	0.71	0.62	0.41	0.71	0.59	0.35	0.56
	Pers.	0.66	0.49	0.50	0.69	0.50	0.45	0.73
	Pol.	0.68	0.52	0.39	0.70	0.51	0.46	0.56
	Sports	0.70	0.57	0.47	0.74	0.55	0.45	0.56

Table 10: Full results by model and domain for Urdu Language.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.68	0.49	0.61	0.67	0.48	0.38	0.48
	Food	0.69	0.45	0.74	0.69	0.53	0.53	0.75
	Hist.	0.65	0.47	0.48	0.66	0.43	0.33	0.48
	M&M	0.64	0.43	0.29	0.75	0.47	0.29	0.29
	Nat. Achv.	0.65	0.45	0.26	0.65	0.49	0.25	0.36
	Nature	0.68	0.52	0.67	0.68	0.56	0.60	0.45
	Pers.	0.63	0.44	0.56	0.64	0.45	0.46	0.52
	Pol.	0.63	0.45	0.41	0.66	0.45	0.37	0.40
	Sports	0.64	0.48	0.42	0.64	0.47	0.44	0.63
Gemma3-12B	Cult.	0.69	0.47	0.64	0.68	0.46	0.63	0.70
	Food	0.70	0.44	0.86	0.70	0.49	0.76	0.95
	Hist.	0.67	0.42	0.63	0.67	0.41	0.55	0.75
	M&M	0.66	0.39	0.32	0.80	0.49	0.37	0.35
	Nat. Achv.	0.66	0.44	0.49	0.67	0.46	0.50	0.42
	Nature	0.69	0.46	0.46	0.67	0.47	0.51	0.52
	Pers.	0.64	0.42	0.67	0.66	0.41	0.65	0.67
	Pol.	0.64	0.42	0.51	0.67	0.42	0.55	0.50
	Sports	0.65	0.42	0.48	0.66	0.42	0.52	0.47
Gemma3-27B	Cult.	0.69	0.47	0.70	0.69	0.50	0.72	0.69
	Food	0.70	0.51	0.88	0.72	0.54	0.87	0.90
	Hist.	0.66	0.41	0.67	0.70	0.47	0.64	0.71
	M&M	0.64	0.42	0.36	0.78	0.48	0.31	0.39
	Nat. Achv.	0.66	0.44	0.51	0.68	0.48	0.53	0.75
	Nature	0.68	0.45	0.61	0.69	0.51	0.63	0.75
	Pers.	0.63	0.42	0.82	0.67	0.45	0.80	0.75
	Pol.	0.63	0.39	0.53	0.68	0.45	0.51	0.58
	Sports	0.66	0.44	0.65	0.66	0.44	0.65	0.80
Qwen2.5-VL-7B	Cult.	0.68	0.38	0.48	0.66	0.34	0.45	0.51
	Food	0.69	0.34	0.61	0.64	0.33	0.51	0.67
	Hist.	0.69	0.44	0.57	0.65	0.34	0.44	0.61
	M&M	0.68	0.34	0.21	0.78	0.41	0.26	0.31
	Nat. Achv.	0.67	0.30	0.66	0.69	0.44	0.68	0.55
	Nature	0.68	0.36	0.40	0.55	0.31	0.66	0.35
	Pers.	0.64	0.36	0.36	0.68	0.45	0.42	0.52
	Pol.	0.65	0.38	0.38	0.69	0.45	0.32	0.51
	Sports	0.68	0.35	0.45	0.70	0.44	0.39	0.53
Qwen3-VL-8B	Cult.	0.67	0.43	0.59	0.66	0.34	0.59	0.70
	Food	0.68	0.40	0.72	0.69	0.42	0.73	0.74
	Hist.	0.67	0.46	0.71	0.67	0.38	0.71	0.65
	M&M	0.67	0.35	0.36	0.69	0.37	0.37	0.28
	Nat. Achv.	0.66	0.41	0.67	0.65	0.35	0.59	0.57
	Nature	0.68	0.43	0.53	0.65	0.34	0.31	0.48
	Pers.	0.63	0.42	0.52	0.65	0.43	0.54	0.52
	Pol.	0.64	0.45	0.45	0.66	0.42	0.47	0.46
	Sports	0.65	0.47	0.48	0.64	0.38	0.53	0.43
GPT-4.1-mini	Cult.	0.70	0.46	0.64	0.70	0.50	0.59	0.75
	Food	0.72	0.45	0.84	0.73	0.54	0.73	0.92
	Hist.	0.69	0.44	0.56	0.70	0.50	0.54	0.68
	M&M	0.68	0.37	0.45	0.90	0.56	0.41	0.45
	Nat. Achv.	0.69	0.42	0.61	0.69	0.52	0.51	0.85
	Nature	0.69	0.41	0.48	0.71	0.51	0.51	0.62
	Pers.	0.65	0.36	0.59	0.69	0.52	0.53	0.73
	Pol.	0.65	0.37	0.49	0.69	0.45	0.43	0.56
	Sports	0.66	0.43	0.52	0.71	0.45	0.53	0.71

Table 11: Full results by model and domain for Barishal Dialect.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.67	0.45	0.65	0.68	0.45	0.42	0.62
	Food	0.69	0.48	0.75	0.69	0.51	0.62	0.77
	Hist.	0.65	0.40	0.47	0.68	0.45	0.33	0.41
	M&M	0.65	0.42	0.25	0.78	0.51	0.25	0.30
	Nat. Achv.	0.65	0.44	0.28	0.67	0.43	0.27	0.31
	Nature	0.69	0.51	0.63	0.69	0.53	0.57	0.49
	Pers.	0.63	0.40	0.52	0.66	0.42	0.44	0.51
	Pol.	0.63	0.41	0.38	0.67	0.41	0.35	0.53
	Sports	0.63	0.45	0.45	0.65	0.45	0.27	0.61
Gemma3-12B	Cult.	0.68	0.45	0.65	0.68	0.45	0.61	0.69
	Food	0.70	0.48	0.85	0.71	0.51	0.71	0.83
	Hist.	0.67	0.40	0.62	0.68	0.40	0.59	0.74
	M&M	0.66	0.42	0.32	0.83	0.51	0.37	0.28
	Nat. Achv.	0.66	0.47	0.54	0.67	0.45	0.49	0.62
	Nature	0.68	0.43	0.51	0.68	0.46	0.45	0.59
	Pers.	0.64	0.41	0.69	0.66	0.41	0.64	0.83
	Pol.	0.64	0.41	0.47	0.68	0.40	0.49	0.54
	Sports	0.65	0.45	0.52	0.66	0.45	0.42	0.50
Gemma3-27B	Cult.	0.68	0.40	0.68	0.69	0.47	0.73	0.77
	Food	0.70	0.44	0.90	0.71	0.50	0.88	0.85
	Hist.	0.66	0.38	0.67	0.69	0.46	0.67	0.77
	M&M	0.66	0.35	0.34	0.79	0.50	0.30	0.34
	Nat. Achv.	0.66	0.45	0.52	0.68	0.43	0.46	0.69
	Nature	0.69	0.43	0.59	0.70	0.50	0.57	0.70
	Pers.	0.63	0.35	0.78	0.67	0.46	0.82	0.80
	Pol.	0.64	0.36	0.55	0.68	0.46	0.48	0.48
	Sports	0.66	0.38	0.66	0.67	0.41	0.66	0.71
Qwen2.5-VL-7B	Cult.	0.67	0.36	0.55	0.63	0.34	0.49	0.60
	Food	0.69	0.31	0.68	0.65	0.29	0.49	0.66
	Hist.	0.69	0.42	0.62	0.67	0.37	0.57	0.50
	M&M	0.67	0.35	0.18	0.79	0.45	0.19	0.18
	Nat. Achv.	0.67	0.34	0.66	0.70	0.39	0.60	0.43
	Nature	0.68	0.37	0.51	0.54	0.28	0.38	0.37
	Pers.	0.63	0.34	0.36	0.67	0.41	0.44	0.54
	Pol.	0.65	0.39	0.36	0.69	0.44	0.31	0.29
	Sports	0.67	0.37	0.52	0.69	0.45	0.43	0.58
Qwen3-VL-8B	Cult.	0.67	0.45	0.60	0.67	0.42	0.57	0.67
	Food	0.68	0.40	0.69	0.69	0.43	0.67	0.74
	Hist.	0.67	0.44	0.68	0.67	0.41	0.68	0.69
	M&M	0.66	0.36	0.29	0.68	0.34	0.25	0.28
	Nat. Achv.	0.65	0.41	0.59	0.66	0.42	0.53	0.59
	Nature	0.67	0.44	0.57	0.64	0.33	0.48	0.45
	Pers.	0.63	0.43	0.57	0.64	0.41	0.46	0.50
	Pol.	0.64	0.44	0.48	0.66	0.41	0.41	0.38
	Sports	0.64	0.42	0.54	0.66	0.39	0.44	0.56
GPT-4.1-mini	Cult.	0.69	0.41	0.70	0.70	0.45	0.58	0.73
	Food	0.72	0.46	0.84	0.74	0.52	0.75	0.91
	Hist.	0.69	0.41	0.64	0.70	0.46	0.64	0.66
	M&M	0.68	0.42	0.48	0.89	0.61	0.42	0.43
	Nat. Achv.	0.69	0.46	0.59	0.70	0.54	0.57	0.80
	Nature	0.70	0.39	0.58	0.71	0.46	0.44	0.57
	Pers.	0.65	0.34	0.67	0.69	0.52	0.54	0.78
	Pol.	0.66	0.36	0.50	0.68	0.45	0.46	0.59
	Sports	0.67	0.41	0.50	0.70	0.48	0.44	0.66

Table 12: Full results by model and domain for Chittagong Dialect.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.69	0.48	0.65	0.68	0.49	0.49	0.51
	Food	0.69	0.45	0.74	0.69	0.50	0.61	0.72
	Hist.	0.66	0.40	0.46	0.68	0.42	0.37	0.51
	M&M	0.65	0.41	0.23	0.79	0.48	0.28	0.31
	Nat. Achv.	0.66	0.49	0.31	0.68	0.45	0.20	0.43
	Nature	0.70	0.56	0.63	0.69	0.56	0.69	0.49
	Pers.	0.64	0.45	0.51	0.68	0.47	0.50	0.53
	Pol.	0.65	0.45	0.40	0.68	0.43	0.41	0.46
	Sports	0.64	0.42	0.50	0.66	0.46	0.45	0.45
Gemma3-12B	Cult.	0.69	0.42	0.67	0.68	0.44	0.65	0.70
	Food	0.71	0.44	0.86	0.71	0.49	0.77	0.88
	Hist.	0.67	0.37	0.63	0.68	0.42	0.60	0.62
	M&M	0.67	0.37	0.35	0.75	0.54	0.34	0.36
	Nat. Achv.	0.66	0.42	0.55	0.68	0.45	0.44	0.44
	Nature	0.69	0.43	0.48	0.68	0.45	0.46	0.60
	Pers.	0.64	0.41	0.68	0.65	0.41	0.67	0.69
	Pol.	0.65	0.40	0.54	0.67	0.41	0.56	0.60
	Sports	0.65	0.40	0.53	0.65	0.42	0.52	0.82
Gemma3-27B	Cult.	0.70	0.47	0.73	0.69	0.48	0.69	0.79
	Food	0.71	0.47	0.90	0.72	0.52	0.91	0.90
	Hist.	0.68	0.44	0.66	0.69	0.45	0.69	0.70
	M&M	0.68	0.41	0.35	0.80	0.51	0.32	0.46
	Nat. Achv.	0.67	0.47	0.48	0.67	0.43	0.53	0.85
	Nature	0.69	0.47	0.62	0.70	0.51	0.68	0.67
	Pers.	0.65	0.39	0.80	0.68	0.46	0.85	0.83
	Pol.	0.65	0.39	0.51	0.69	0.46	0.47	0.50
	Sports	0.67	0.43	0.65	0.66	0.42	0.58	0.67
Qwen2.5-VL-7B	Cult.	0.68	0.37	0.51	0.69	0.36	0.49	0.53
	Food	0.69	0.33	0.66	0.71	0.41	0.54	0.68
	Hist.	0.69	0.42	0.60	0.70	0.44	0.55	0.60
	M&M	0.68	0.35	0.21	0.83	0.53	0.29	0.23
	Nat. Achv.	0.67	0.41	0.69	0.70	0.42	0.48	0.56
	Nature	0.68	0.41	0.37	0.71	0.44	0.58	0.34
	Pers.	0.64	0.36	0.46	0.67	0.41	0.44	0.43
	Pol.	0.65	0.39	0.39	0.68	0.44	0.36	0.30
	Sports	0.67	0.40	0.44	0.72	0.49	0.42	0.41
Qwen3-VL-8B	Cult.	0.67	0.45	0.65	0.67	0.35	0.62	0.64
	Food	0.69	0.40	0.73	0.68	0.40	0.69	0.74
	Hist.	0.68	0.43	0.69	0.66	0.35	0.63	0.72
	M&M	0.67	0.36	0.38	0.69	0.37	0.30	0.35
	Nat. Achv.	0.67	0.46	0.63	0.66	0.42	0.57	0.81
	Nature	0.68	0.43	0.38	0.64	0.30	0.35	0.43
	Pers.	0.64	0.42	0.50	0.65	0.42	0.44	0.46
	Pol.	0.64	0.42	0.43	0.66	0.39	0.44	0.34
	Sports	0.66	0.47	0.50	0.66	0.38	0.47	0.58
GPT-4.1-mini	Cult.	0.70	0.44	0.65	0.71	0.52	0.58	0.72
	Food	0.71	0.41	0.85	0.73	0.51	0.73	0.88
	Hist.	0.69	0.44	0.61	0.71	0.50	0.57	0.68
	M&M	0.68	0.36	0.48	0.90	0.64	0.38	0.49
	Nat. Achv.	0.69	0.43	0.62	0.71	0.53	0.62	0.82
	Nature	0.70	0.39	0.49	0.71	0.50	0.51	0.59
	Pers.	0.65	0.37	0.57	0.70	0.51	0.57	0.74
	Pol.	0.67	0.38	0.42	0.69	0.49	0.51	0.57
	Sports	0.68	0.42	0.53	0.69	0.48	0.50	0.61

Table 13: Full results by model and domain for Noakhali Dialect.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.68	0.49	0.63	0.68	0.49	0.50	0.59
	Food	0.70	0.50	0.73	0.70	0.52	0.56	0.72
	Hist.	0.67	0.48	0.44	0.68	0.44	0.42	0.40
	M&M	0.66	0.44	0.27	0.80	0.52	0.29	0.30
	Nat. Achv.	0.67	0.52	0.35	0.66	0.44	0.27	0.45
	Nature	0.69	0.55	0.67	0.69	0.57	0.69	0.41
	Pers.	0.63	0.42	0.59	0.67	0.46	0.55	0.53
	Pol.	0.64	0.41	0.42	0.69	0.45	0.38	0.41
	Sports	0.65	0.48	0.48	0.66	0.47	0.37	0.46
Gemma3-12B	Cult.	0.69	0.45	0.62	0.68	0.44	0.62	0.82
	Food	0.70	0.47	0.85	0.71	0.51	0.78	0.82
	Hist.	0.67	0.43	0.64	0.68	0.44	0.60	0.76
	M&M	0.65	0.41	0.37	0.82	0.51	0.32	0.44
	Nat. Achv.	0.65	0.45	0.52	0.68	0.46	0.52	0.71
	Nature	0.68	0.43	0.44	0.69	0.46	0.54	0.58
	Pers.	0.63	0.42	0.66	0.65	0.40	0.71	0.76
	Pol.	0.65	0.41	0.55	0.67	0.40	0.54	0.62
	Sports	0.64	0.44	0.50	0.66	0.44	0.52	0.54
Gemma3-27B	Cult.	0.70	0.52	0.69	0.68	0.49	0.72	0.64
	Food	0.71	0.50	0.88	0.72	0.52	0.90	0.87
	Hist.	0.68	0.47	0.69	0.69	0.47	0.66	0.75
	M&M	0.67	0.43	0.33	0.84	0.52	0.34	0.43
	Nat. Achv.	0.68	0.48	0.52	0.69	0.43	0.55	0.73
	Nature	0.69	0.49	0.59	0.70	0.54	0.60	0.67
	Pers.	0.64	0.42	0.78	0.69	0.48	0.78	0.79
	Pol.	0.65	0.43	0.57	0.68	0.46	0.51	0.51
	Sports	0.67	0.48	0.66	0.68	0.45	0.53	0.84
Qwen2.5-VL-7B	Cult.	0.69	0.49	0.55	0.69	0.47	0.40	0.59
	Food	0.69	0.43	0.68	0.72	0.46	0.57	0.68
	Hist.	0.70	0.51	0.60	0.71	0.52	0.51	0.50
	M&M	0.69	0.44	0.22	0.81	0.57	0.26	0.20
	Nat. Achv.	0.68	0.49	0.69	0.69	0.50	0.71	0.51
	Nature	0.70	0.51	0.42	0.71	0.50	0.70	0.40
	Pers.	0.65	0.45	0.47	0.68	0.47	0.39	0.53
	Pol.	0.66	0.47	0.39	0.67	0.49	0.36	0.33
	Sports	0.69	0.49	0.48	0.72	0.54	0.44	0.44
Qwen3-VL-8B	Cult.	0.66	0.43	0.57	0.66	0.35	0.65	0.66
	Food	0.68	0.41	0.75	0.68	0.41	0.70	0.73
	Hist.	0.67	0.44	0.70	0.67	0.38	0.72	0.65
	M&M	0.66	0.34	0.32	0.67	0.38	0.27	0.38
	Nat. Achv.	0.65	0.41	0.66	0.64	0.42	0.55	0.65
	Nature	0.67	0.42	0.57	0.65	0.34	0.50	0.49
	Pers.	0.62	0.41	0.57	0.64	0.44	0.53	0.54
	Pol.	0.64	0.42	0.45	0.67	0.45	0.47	0.44
	Sports	0.64	0.44	0.48	0.65	0.36	0.42	0.61
GPT-4.1-mini	Cult.	0.69	0.42	0.66	0.70	0.50	0.58	0.78
	Food	0.72	0.44	0.86	0.73	0.54	0.68	0.89
	Hist.	0.69	0.45	0.63	0.69	0.48	0.62	0.72
	M&M	0.69	0.39	0.53	0.91	0.63	0.39	0.43
	Nat. Achv.	0.69	0.45	0.57	0.70	0.52	0.58	0.74
	Nature	0.69	0.41	0.55	0.71	0.53	0.49	0.69
	Pers.	0.64	0.39	0.59	0.70	0.51	0.46	0.72
	Pol.	0.66	0.35	0.46	0.69	0.46	0.46	0.51
	Sports	0.67	0.43	0.53	0.71	0.53	0.50	0.74

Table 14: Full results by model and domain for Rangpur Dialect.

Models	Domain	Zero-Shot Caption B-F1	Zero-Shot Caption LLM	Zero-Shot VQA Accuracy (%)	Few-Shot Caption B-F1	Few-Shot Caption LLM	Few-Shot VQA Accuracy (%)	CoT VQA Accuracy (%)
Gemma3-4B	Cult.	0.67	0.47	0.60	0.68	0.45	0.45	0.58
	Food	0.70	0.46	0.72	0.70	0.50	0.62	0.64
	Hist.	0.66	0.42	0.44	0.68	0.43	0.40	0.47
	M&M	0.65	0.42	0.27	0.80	0.50	0.35	0.36
	Nat. Achv.	0.65	0.47	0.31	0.68	0.48	0.29	0.45
	Nature	0.69	0.58	0.59	0.69	0.56	0.63	0.52
	Pers.	0.63	0.45	0.58	0.67	0.45	0.54	0.51
	Pol.	0.64	0.45	0.40	0.68	0.43	0.42	0.40
	Sports	0.64	0.47	0.47	0.67	0.47	0.42	0.44
Gemma3-12B	Cult.	0.68	0.40	0.62	0.68	0.43	0.65	0.78
	Food	0.70	0.40	0.82	0.71	0.48	0.67	0.91
	Hist.	0.66	0.37	0.64	0.69	0.39	0.64	0.64
	M&M	0.66	0.39	0.35	0.80	0.49	0.33	0.45
	Nat. Achv.	0.65	0.38	0.46	0.67	0.43	0.38	0.58
	Nature	0.68	0.41	0.47	0.68	0.45	0.54	0.42
	Pers.	0.63	0.31	0.68	0.65	0.34	0.65	0.81
	Pol.	0.63	0.33	0.54	0.67	0.39	0.52	0.59
	Sports	0.65	0.39	0.45	0.66	0.40	0.44	0.50
Gemma3-27B	Cult.	0.69	0.48	0.70	0.69	0.48	0.68	0.78
	Food	0.71	0.45	0.84	0.72	0.53	0.82	0.86
	Hist.	0.69	0.44	0.66	0.69	0.46	0.63	0.75
	M&M	0.67	0.42	0.34	0.84	0.51	0.33	0.45
	Nat. Achv.	0.68	0.43	0.51	0.68	0.45	0.49	0.69
	Nature	0.69	0.46	0.59	0.70	0.50	0.60	0.63
	Pers.	0.65	0.36	0.79	0.69	0.48	0.82	0.78
	Pol.	0.66	0.38	0.51	0.68	0.46	0.47	0.55
	Sports	0.68	0.42	0.63	0.67	0.41	0.56	0.68
Qwen2.5-VL-7B	Cult.	0.67	0.35	0.52	0.69	0.32	0.42	0.53
	Food	0.69	0.32	0.55	0.71	0.35	0.45	0.59
	Hist.	0.68	0.42	0.58	0.70	0.43	0.55	0.52
	M&M	0.67	0.35	0.21	0.83	0.53	0.22	0.22
	Nat. Achv.	0.67	0.36	0.66	0.70	0.40	0.71	0.52
	Nature	0.69	0.39	0.43	0.70	0.44	0.60	0.38
	Pers.	0.64	0.36	0.49	0.67	0.41	0.48	0.51
	Pol.	0.66	0.38	0.39	0.69	0.44	0.34	0.36
	Sports	0.67	0.41	0.45	0.69	0.44	0.37	0.48
Qwen3-VL-8B	Cult.	0.66	0.45	0.66	0.65	0.35	0.59	0.68
	Food	0.68	0.44	0.71	0.69	0.45	0.68	0.73
	Hist.	0.67	0.44	0.72	0.66	0.35	0.67	0.72
	M&M	0.66	0.34	0.42	0.67	0.37	0.44	0.37
	Nat. Achv.	0.67	0.45	0.66	0.65	0.34	0.69	0.72
	Nature	0.68	0.47	0.48	0.64	0.32	0.42	0.44
	Pers.	0.63	0.41	0.49	0.64	0.45	0.56	0.56
	Pol.	0.64	0.44	0.42	0.66	0.42	0.41	0.42
	Sports	0.66	0.44	0.53	0.64	0.39	0.51	0.59
GPT-4.1-mini	Cult.	0.69	0.43	0.63	0.70	0.51	0.59	0.71
	Food	0.72	0.46	0.84	0.72	0.50	0.66	0.86
	Hist.	0.69	0.46	0.58	0.70	0.49	0.58	0.68
	M&M	0.69	0.40	0.47	0.90	0.59	0.43	0.50
	Nat. Achv.	0.69	0.48	0.64	0.70	0.55	0.62	0.76
	Nature	0.70	0.45	0.53	0.72	0.54	0.47	0.67
	Pers.	0.65	0.36	0.63	0.70	0.51	0.51	0.71
	Pol.	0.66	0.40	0.48	0.68	0.47	0.44	0.63
	Sports	0.68	0.46	0.47	0.70	0.50	0.47	0.73

Table 15: Full results by model and domain for Sylhet Dialect.