# ChitroJera: A Regionally Relevant Visual Question Answering Dataset for Bangla

Deeparghya Dutta Barua[1⋆], Md Sakib Ul Rahman Sourove[1⋆], Md Fahim[1,2⋆], Fabiha Haider[1], Fariha Tanjim Shifat[1], Md Tasmim Rahman Adib[1], Anam Borhan Uddin[1], Md Farhan Ishmam[1,3], and Md Farhad Alam[1](✉)

[1] Research and Development, Penta Global Limited
[2] CCDS Lab, Independent University, Bangladesh
[3] Kahlert School of Computing, University of Utah
{deeparghya.csedu, farhan.ishmam, pdcsedu}@gmail.com

**Abstract** Visual Question Answer (VQA) poses the problem of answering a natural language question about a visual context. Bangla, despite being a widely spoken language, is considered low-resource in the realm of VQA due to the lack of proper benchmarks, challenging models known to be performant in other languages. Furthermore, existing Bangla VQA datasets offer little regional relevance and are largely adapted from their foreign counterparts. To address these challenges, we introduce a large-scale Bangla VQA dataset, ChitroJera, totaling over 15k samples from diverse and locally relevant data sources. We assess the performance of text encoders, image encoders, multimodal models, and our novel dual-encoder models. The experiments reveal that the pre-trained dual-encoders outperform other models of their scale. We also evaluate the performance of current large vision language models (LVLMs) using prompt-based techniques, achieving the overall best performance. Given the underdeveloped state of existing datasets, we envision ChitroJera expanding the scope of Vision-Language tasks in Bangla. Our code and data are available at: http://github.com/farhanishmam/ChitroJera.

**Keywords:** Visual Question Answering · Low Resource Languages · Multimodal Models.

## 1 Introduction & Related Work

Visual Question Answering (VQA) has gained relevance lately with the onset of transformer-based models, facilitating a better understanding of language and context in different modalities [25]. This has led to the focus in VQA research shifting from the perception of language and vision to understanding the reasoning of these opaque systems [46]. VQA systems are also being used to aid in visual impairment [19], enhance robotic systems [47], expedite the screening of medical conditions from relevant imagery [36], and so on. The performance of these systems is strongly coupled with the quality of the datasets they are

⋆ Equal Contribution

Table 1: **Comparison between existing datasets** based on the number of questions (#Q), answers (#A), and images (#I), source of images (Img Src.), annotation (Annot.) and Validation (Val.) methods, Question Type (QT), and categorical metadata availability (CMA). gTrans means Google Translate.

| Datasets | #Q | #A | #I | Img Src. | Annot. | Val. | QT | CMA |
|---|---|---|---|---|---|---|---|---|
| Bengali VQA v1 | 5000 | 2 | 500 | English | gTrans | Authors | Binary | ✗ |
| Bengali CLEVR | 12291 | 1600 | 1271 | English | gTrans | Authors | WH | ✗ |
| Bengali VQA 2.0 | 13046 | 2 | 3280 | English | Manual | N/A | Binary | ✗ |
| CVQA (Bengali subset) | 286 | 780 | 136 | Bangla | Manual | Natives | WH | ✓ |
| ChitroJera (Ours) | 15292 | 5542 | 15147 | Bangla | GPT-4 Turbo | Experts | WH | ✓ |

trained on [18, 41]. With that in mind, VQA datasets in English are being designed with increasing complexity, leaning more towards advanced reasoning instead of simple answers [40, 46]. However, the landscape for most low-resource languages speaks differently, with the major pain point being the lack of substantial datasets that can address even the most basic of answers.

Despite being a language with around 284 million[4] speakers, Bangla has received limited exposure to the domain of visual question answering. The issues are multifaceted, with one aspect being the lack of pre-trained vision language models (VLMs) to entertain the idiosyncrasies of the language, and the other being the lack of datasets tailored for this particular purpose, further compounding the issue of VLM unavailability.

Presently, only four instances of VQA datasets have been compiled for Bangla [26, 43, 45], all of which have certain limitations as per Tab. 1. The first work presents two datasets, Bengali-VQA-v1 and Bengali CLEVR, built on top of the existing English VQA datasets — VQA v1 [4] and CLEVR [27] respectively. The second work presents Bengali-VQA-2.0, compiled from the English VQA v2.0 dataset. Among these, Bengali-VQA-v1 and Bengali-VQA-2.0 offer limited applicability to non-trivial VQA tasks, as the questions are restricted to binary answers. While the answers in Bengali CLEVR encompass multiple classes, the issue with all these datasets is that the source images and texts are derived from English datasets, which lack the geographical context associated with Bangla. Also, Bengali VQA 2.0 magnifies its scale by relying on tactics that yield repetitive samples with minor differences. Lastly, CVQA [45] addresses some of these issues, but as it explores 26 languages, the focus on Bangla is quite thin, yielding only 286 samples.

With the intent of addressing the shortcomings within the existing space of Bangla VQA, our contributions are as follows:

**ChitroJera Dataset**: We propose a new dataset for Bangla VQA, named ChitroJera, comprising 15k images and questions, synthesized using OpenAI

[4] https://www.ethnologue.com/insights/ethnologue200/ (Current as of writing; figures are subject to change.)

GPT-4 Turbo with curated prompting and validated by linguistic experts. The images and text have a Bangla regional flavor, *i.e.* that they capture the connotations associated with the Bangla-speaking region. We ensure diversity by imposing restrictions on the number of questions per image. For better analysis, we provide a categorical breakdown of the samples based on the subject of the questions. Comparison between ours and existing datasets is given in Table 1.

**Dual Encoder Models**: Due to the lack of VLMs that are aligned on Bangla text with regional images, we introduce novel dual encoder-based models that outperform existing unimodal models and VLMs trained on English data sources, showing promising performance at its scale.

**Extensive Experimentation**: We conduct experiments on our dataset using state-of-the-art and widely adopted text encoders, visual encoders, and multimodal models fine-tuned for this task. We observe the performance of dual encoder models with and without pretraining, with ablations on different pretraining objectives and batch sizes. Finally, we assess both open and closed-source LVLMs via zero-shot prompting and compare them. Experiments demonstrate that the dual encoder-based models we propose outperform all open-source LVLMs in zero-shot evaluations across all metrics.

## 2 ChitroJera Dataset

The ChitroJera dataset is meticulously curated and annotated for regionally relevant Bangla VQA. We source the data from internet content representative of the Bengali communities.

### 2.1 Dataset Collection

The question-answer pairs in the dataset have been collected from existing Bangla images and image captions found in the BanglaLekhaImageCaptions [44], Bornon [48], and BNATURE [17] datasets. These sources contain images from the internet, which have been collected using keywords relevant to Bangladesh and its vicinity, transitively ensuring that our dataset is regionally relevant. A detailed breakdown of the source distribution is given in Table 2.

### 2.2 Data Preprocessing

We corrected image-caption mismatches, deduplicated images, and removed erroneous ICC color profiles. A single image can have multiple captions. For images with more than 3 captions, we choose the longest and the shortest captions, and a third caption having the highest BERTScore with the former two. The reason behind this choice is to capture a diverse depiction of the image. While shorter captions typically convey the broader context, longer captions tend to include finer details. These three (or fewer) captions are concatenated to form the textual ground truth context for the question generation.

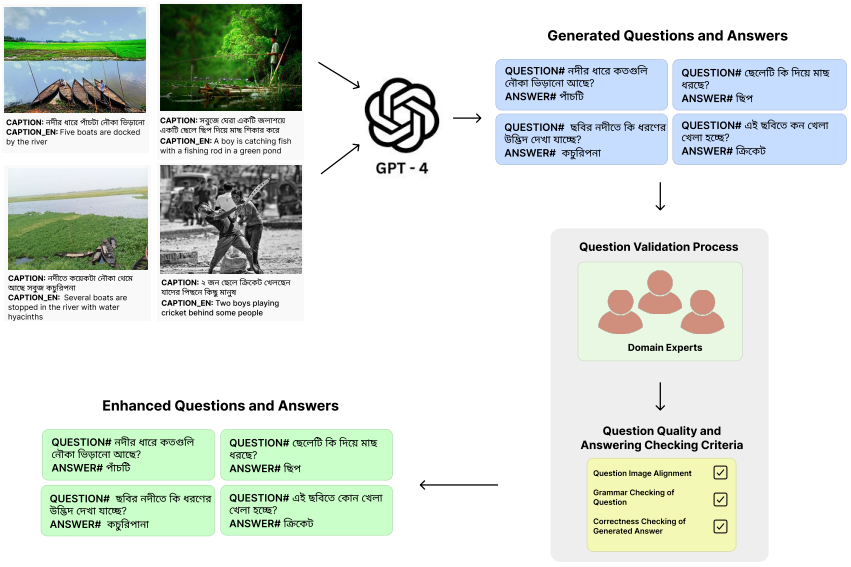



Figure 1: **Overview of the dataset generation pipeline.** The image-caption pairs are passed to GPT-4 Turbo using a curated prompt to generate QA pairs, then validated and corrected by the domain experts.

### 2.3 QA Pair Generation

We compared QA pairs generated by GPT-4 Turbo and Google Gemini 1.5 Pro over 1,000 image–caption pairs, and relied on two domain experts to choose the superior output for each instance. Since the evaluation criteria were binary and there were exactly two annotators, we measured inter-annotator agreement using Cohen's $\kappa$ between the annotators, obtaining a score of 89.27%. GPT-4 Turbo was preferred in most of the cases, and is therefore used for question synthesis. We prompt the model in such a manner to generate complex and diverse questions while constraining answers to one to three words, ensuring the response remains concise and minimizes irrelevant information. The prompt used for QA pair generation has been reported in Sec. 8.

### 2.4 Dataset Annotation

We verify the quality and consistency of the synthesized QA pairs by defining a few evaluation criteria, namely caption-question alignment, image-question alignment, question correctness, and answer correctness. Our approach to data validation is methodological. We employ two experts in Bangla linguistics to evaluate the dataset on the aforementioned criteria over a subset of 2500 randomly selected samples due to resource constraints. Both experts are native Bangla speakers, accredited in linguistic proficiency, possess strong cultural

Table 2: Dataset statistics of ChitroJera.

| **Source Distribution** | |
|---|---|
| BanglaLekhaImageCaptions | 8600 |
| Bornon | 4292 |
| BNATURE | 2400 |
| **Splits** | |
| Train | 12231 |
| Validation | 1529 |
| Test | 1532 |
| **General Statistics** | |
| Samples | 15292 |
| Captions | 14927 |
| Images | 15147 |
| Questions | 13299 |
| Answers | 5542 |
| WH-words | 11 |
| Categories/Types | 17 |

| **QA Statistics** | **Q** | **A** |
|---|---|---|
| Mean character length | 33.50 | 7.10 |
| Max character length | 105 | 45 |
| Min character length | 11 | 1 |
| Mean word count | 5.86 | 1.43 |
| Max word count | 17 | 4 |
| Min word count | 3 | 1 |
| **QA Pair LLM Selection** | | |
| #Samples | | 1000 |
| Cohen's $\kappa$ coefficient | | 89.27% |
| **QA Pair Validation** | | |
| #Samples | | 2500 |
| **Accuracy** | **#** | **%** |
| Annotator 1 | 2462 | 98.48 |
| Annotator 2 | 2476 | 99.04 |

awareness, and conform to the same principles in assessing grammatical correctness. We hired them on an hourly basis. The rationale for hiring domain experts instead of general annotators lies in their deeper understanding of the linguistic nuances of Bangla, enabling them to uphold the quality of the dataset. Since the overhead of cross-checking between two annotators is low, any inter-annotator disagreement is disputed directly via discussion. If no consensus is reached, we discard the ambiguous sample. However, such cases are rare, with only 5 samples being flagged. The annotation statistics are shown in Tab. 2.

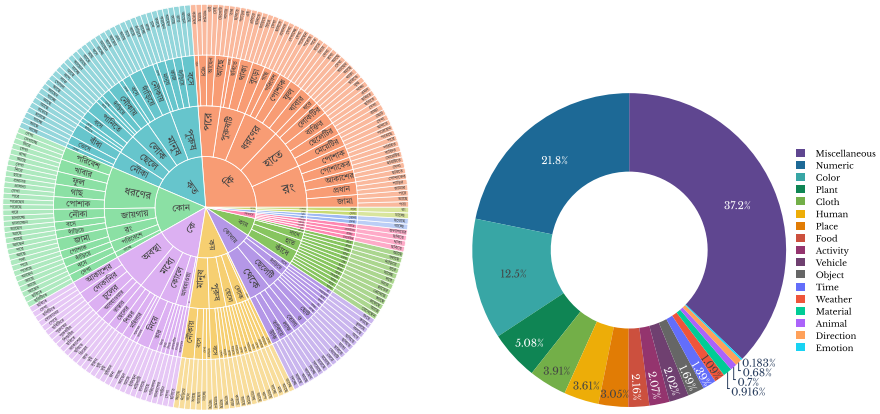
### 2.5 Dataset Statistics

We randomly split the dataset into training, validation, and test sets with a ratio of 80:10:10. To capture a more diverse representation of different contexts, we restrict the number of questions for each image to one or, at most, two. The number of unique questions is lesser than the total number of question, i.e., the same question has been asked on different images. This enables the models to generalize effectively, even when the visual context significantly differs. While the ratio of unique questions to unique answers is approximately 2.4, it does not necessarily indicate a sparsity in the training distribution. This is due to the inclusion of suffixes in the answers, i.e., the same root word can take multiple forms. The statistics of ChitroJera are shown in Tab. 2.

**Question Statistics** The questions generated by GPT-4 Turbo range from a minimum of 3 words to a maximum of 17 words, with a mean of 6 words. To

assess the diversity of the questions, we chose a set of 11 wh-question words in Bangla that comprehensively cover all the grammatically correct questions. These are — “কি” (what/tag questions), “কোন” (which), “কত” (how many/much), “কোথায়” (where), “কয়” (how many/much), “কে” (who), “কার” (singular whose), “কখন” (when), “কিভাবে” (how), “কাদের” (plural whose) and “কবে” (when).

Figure 2a offers an overview of the distribution of the keywords, respectively. We hypothesize that the low sample count of questions using “কার” (singular whose), “কখন” (when), “কাদের” (plural whose), and “কবে” (when) is due to the difficulty in assessing temporal and possessive qualities from still images and captions that are unassuming of their subjects.




(a) (b)

Figure 2: (a) Question keyword and (b) Answer category distribution of Chitro-Jera dataset.

**Answer Statistics** The prompt used to generate the QA pairs limits the answers to short phrases, with most being single-word answers. For gauging the distribution of the answers in terms of content, we employ a keyword-based method to classify them into the discrete and exclusive categories of “Numeric”, “Food”, “Place”, “Weather”, “Animal”, “Color”, “Plant”, “Material”, “Activity”, Emotion”, “Cloth”, “Direction”, “Human”, “Vehicle”, “Time” and “Object”; outlined in Figure 2b.

## 3 Baselines

We evaluate image-only, text-only, and multimodal (image + text) settings, and compare the performance of LLMs and VLMs. For evaluation, we consider exact match accuracy, BERTScore [54], and LAVE score [39].

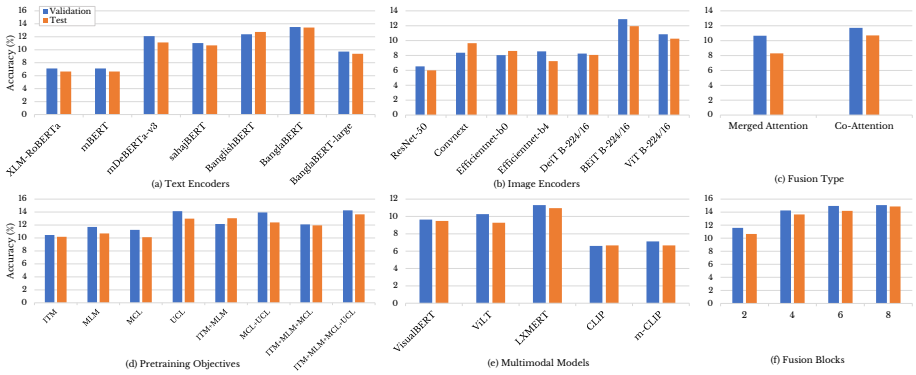



Figure 3: Validation and test accuracy across different text encoders, image encoders, multimodal models, fusion types, number of fusion blocks, and pretraining objectives.

### 3.1 Fine-tuning

We evaluate the dataset on models fine-tuned with our train split. The problem is treated as a multi-class classification task, following other close-ended VQA approaches [40,46]. Unimodal fine-tuning is performed on text and image modalities separately. In this case, the text models involve multilingual models like XLM-Roberta [11], mBERT [35], and mDeBERTa-v3 [23], and Bangla language models such as sahajBERT[5], BanglishBERT and BanglaBERT [8], where BanglaBERT achieves the best performance (Fig. 3a). Here, the input, $x_t$ = *Ques:* `{question}` `[SEP]` *Caption:* `{caption}`. For the visual modality, we run our experiments on ResNet [22], Convnext [38], Efficientnet, [50], DeiT [51], BEiT, [6], and ViT [14], where BEiT and ViT outperform the others (Fig. 3b). We also fine-tune multimodal models (VLMs), such as VisualBERT [33], ViLT [29], LXMERT [49], CLIP [42], and m-CLIP [9], where LXMERT has the best standing. The formal representation of the [CLS] token (denoted as $\mathbf{h}$) extraction is as follows:

$$\text{Text encoders:} \quad \mathbf{h} = f_T(x_t) \quad \text{and} \quad \text{image encoders:} \quad \mathbf{h} = f_I(x_{img}).$$

$$\text{For VisualBERT and ViLT:} \quad \mathbf{h} = f_{VL}(x_t, x_{\text{img}}).$$

$$\text{For LXMERT, CLIP, m-CLIP:} \begin{cases} \mathbf{h}_T, \mathbf{h}_I = f_{VL}(x_t, x_{\text{img}}) \\ \mathbf{h} = [\mathbf{h}_T; \mathbf{h}_I]. \end{cases}$$

If there are multiple [CLS] tokens for each modality, we simply concatenate them to have a single representation, $\mathbf{h}$. As per the definition of a multi-class classification problem, $\mathbf{h}$ is then passed to an MLP. To fine-tune the model,

[5] https://github.com/tanmoyio/sahajbert

the system calculates cross-entropy loss by evaluating MLP outputs against the ground truth labels.

### 3.2 Pretrained Dual Encoders

From the discussion in Section 3.1, pre-trained multimodal models are typically trained on image-English text pairs, resulting in a limited semantic understanding of Bangla text. To address this gap, we adopt a dual-encoder approach to extract semantic features. However, these unimodal models initially lack alignment between modalities, which we address by training a multi-modal fusion network on a pretraining dataset with specific pretraining objectives.

Following the experimental results, we utilize the best-performing models from each modality: BanglaBERT for text, and both BEiT and ViT separately for the image. For pretraining datasets, we consider the image-caption dataset to align the dual encoders for both modalities. As described in Section 2.1, we use BanglaLekhaImageCaptions, Bornon, and BNATURE to generate our VQA dataset. Therefore, we omit these datasets for our pretraining tasks and instead use the BanCap [28] dataset.

**Pretraining Mechanism** During the pretraining stage, the image $x_{img}$ and its corresponding captions are fed $x_t$ into image encoder $f_I$ and text encoder $f_T$ separately, to extract image representations $\mathbf{h}_I$ and text representations $\mathbf{h}_T$ where $\mathbf{h}_I = f_I(x_{img}); \mathbf{h}_T = f_T(x_t)$. Here, $\mathbf{h}_I \in \mathbb{R}^{n_v \times d_v}$ and $\mathbf{h}_T \in \mathbb{R}^{n_t \times d_t}$, where $n_v$ and $n_t$ are the number of visual and textual tokens, respectively, and $d_v$ and $d_t$ are their dimensionalities. A fusion module is employed to align the image and text representations. We explore two types of fusion modules: merged attention and co-attention like [15, 24].

**Merged Attention:** Merged attention integrates information from both modalities into a single attention mechanism. We concatenate both representations $\mathbf{h}_{VL} = [\mathbf{h}_I; \mathbf{h}_T] \in \mathbb{R}^{(n_v+n_t)\times d}$. $\mathbf{h}_{VL}$ is passed into $B$ different fusion blocks which is typically the transformer-encoder blocks [52].

**Co-Attention:** In the co-attention module, the features are processed through $D$ separate transformer blocks, utilizing techniques like cross-attention for cross-modal interaction. For each block $i$, the representations are calculated as follows:

$$
\begin{aligned}
Q_I^i, K_I^i, V_I^i = W_{Q_I}^i \mathbf{h}_I^i, W_{K_I}^i \mathbf{h}_I^i, W_{V_I}^i \mathbf{h}_I^i; \quad & Q_T^i, K_T^i, V_T^i = W_{Q_T}^i \mathbf{h}_T^i, W_{K_T}^i \mathbf{h}_T^i, W_{V_T}^i \mathbf{h}_T^i \\
\mathbf{h}_I^i = \text{Self-Attn}(Q_I^i, K_I^i, V_I^i); \quad & \mathbf{h}_T^i = \text{Self-Attn}(Q_T^i, K_T^i, V_T^i) \\
\mathbf{h'}_I^i = \text{Cross-Attn}(Q_I^i, K_T^i, V_T^i); \quad & \mathbf{h'}_T^i = \text{Cross-Attn}(Q_T^i, K_I^i, V_I^i) \\
\mathbf{h}_I^{i+1} = \text{MLP}(\mathbf{h'}_I^i); \quad & \mathbf{h}_T^{i+1} = \text{MLP}(\mathbf{h'}_T^i).
\end{aligned}
$$

**Pretraining Objectives:** We utilize four different losses, namely Masked Language Modeling (MLM) [13, 55], Image-Text Matching (ITM) [32], Multimodal

Contrastive Loss (MCL) [21], and Unimodal Contrastive Loss (UCL) [34], for the pretraining process. These are the most common and widely used pretraining objectives in the multimodal domain. From Figure 3d, combining all pretraining objectives as $\mathcal{L}_{PT} = \mathcal{L}_{ITM} + \mathcal{L}_{MLM} + \mathcal{L}_{MCL} + \mathcal{L}_{UCL}$.

**Fusion Type:** Following Fig. 3c, the Co-Attention fusion outperforms the Merged Attention fusion. Hence, we choose Co-Attention as the type of fusion for further experimentation.

Table 3: Evaluation of dual-encoder pre-training with $\mathcal{L}_{PT}$ training objectives using accuracy and BERTScore.

| **Dual Encoders** | **Performance Metrics** | | | |
|---|---|---|---|---|
| | **Validation** | | **Test** | |
| | **Accuracy** | **BERTScore** | **Accuracy** | **BERTScore** |
| **w/o Pretraining** | | | | |
| BanglaBERT + ViT-224 | 11.45 | 89.01 | 10.64 | 88.71 |
| BanglaBERT + BEiT-224 | **12.68** | **89.87** | **11.11** | **89.26** |
| **with Pretraining** | | | | |
| BanglaBERT + ViT-224 | 14.26 | **90.78** | **13.64** | **90.59** |
| BanglaBERT + BEiT-224 | **14.45** | 89.50 | 13.12 | 89.56 |

**Effect of Pre-training:** Table 3 presents the performance results of the selected dual-encoders with and without pretraining. It indicates that pretraining enhances the dual encoder accuracy by approx. 2-3% on both the validation and test datasets. Given that our pretraining dataset contains only 44k samples, the improvement is modest and is likely to improve with more training.

**Feature Aggregation based Fine-tuning** When employing a co-attention-based network for modality fusion, we get image-aware text representations $\mathbf{h}'_L$ and text-aware image representations $\mathbf{h}'_I$ after pre-training. While fine-tuning for VQA classification, we extract the [CLS] token representations, denoted as ${h'}_T^{[\text{CLS}]}$ and ${h'}_I^{[\text{CLS}]}$. To derive the ultimate representation for classification, we explore two aggregation techniques.

- **Concat-based:** The final representation $z$ is calculated as follows:

$$z = \texttt{MLP}([{h'}_T^{[\text{CLS}]}; {h'}_I^{[\text{CLS}]}]).$$

- **Summed-based:** In this case, the final representation $z$ is calculated using:

$$z = \texttt{MLP}([{h'}_T^{[\text{CLS}]} + {h'}_I^{[\text{CLS}]}]).$$

This aggregated representation $z$ is then fed into the classification head, followed by a linear layer for prediction.

Table 4: The performance of dual-encoder models using different modality aggregation techniques and large language models on the ChitroJera dataset.

| Models | Performance Metrics | | | | | |
|---|---|---|---|---|---|---|
| | Validation | | | Test | | |
| | Acc | BScore | LAVE | Acc | BScore | LAVE |
| **Dual Encoder Fusion Type** | | | | | | |
| BanglaBERT-ViT [Concat] | 14.26 | 90.78 | 15.83 | 13.64 | 90.59 | 16.08 |
| BanglaBERT-BEiT [Concat] | 14.45 | 89.50 | 20.60 | 13.12 | 89.56 | 20.10 |
| BanglaBERT-ViT [Sum] | 14.08 | 89.71 | 18.88 | 13.61 | 89.32 | 23.46 |
| BanglaBERT-BEiT [Sum] | 14.03 | 89.76 | 18.58 | 14.45 | 90.22 | 20.44 |
| **Open Source LLMs [Monolingual]** | | | | | | |
| BLIP-2 | 11.74 | 87.29 | 19.74 | 10.35 | 86.92 | 19.71 |
| InstructBLIP | 5.29 | 70.25 | 6.68 | 5.67 | 71.14 | 7.24 |
| LLaVa-1.5-7B | 7.93 | 74.86 | 8.80 | 6.73 | 73.75 | 7.93 |
| LLaVA-OneVision-7B | 7.89 | 73.64 | 8.79 | 6.68 | 71.64 | 7.88 |
| **Open Source LLMs [Multilingual]** | | | | | | |
| PaliGemma-3B | 8.44 | 79.26 | 9.37 | 8.98 | 80.72 | 10.59 |
| Pangea-7B | 11.26 | 86.28 | 17.94 | 10.28 | 86.26 | 18.88 |
| Qwen2.5-VL-7B | 12.31 | 88.04 | 18.36 | 12.04 | 87.93 | 18.45 |
| Phi-3.5-Vision | 10.67 | 83.57 | 19.01 | 10.31 | 83.26 | 19.97 |
| InternVL2-8B | 11.98 | 87.33 | 17.22 | 11.24 | 87.01 | 16.85 |
| **Closed Weights/Source LLMs** | | | | | | |
| Gemini 2.0 Flash | 23.07 | 90.15 | **62.31** | 26.58 | 89.15 | **66.08** |
| Claude 3.7 Sonnet | 21.55 | 88.72 | 52.76 | 28.09 | 89.48 | 63.82 |
| GPT-4o | 31.58 | 92.01 | 56.28 | 30.22 | 91.79 | 58.54 |
| GPT-4 Turbo | **33.35** | **92.28** | 61.30 | **32.83** | **92.18** | 57.79 |

### 3.3 LLM Prompting

We also investigate the performance of LLMs & V-LLMs using our dataset through zero-shot prompting techniques. For this, we input an image $x_{img}$ along with a question $x_t$ along with the proper instruction and ask the LVLMs to generate the answer based on the image and question. The selection of the models considers a few criteria: whether they are monolingual and multilingual or not, and whether the model is open weights/source or closed weights/source.

The monolingual open weights/source include BLIP-2 [31], InstructBLIP [12], LLaVa-1.5-7B [37], and LLaVa-OneVision-7B [30]. For open weights/source

multilingual models, we chose PaliGemma-3B [7], Pangea-7B [53], Qwen2.5-VL-7B [5], Phi-3.5-Vision [1], and InternVL2-8B [10]. The closed weights/source offerings include commercial models with vision capabilities, namely Gemini 2.0 Flash, Claude 3.7 Sonnet, GPT-4 Turbo, and GPT-4o. The prompts used for our experiments are reported in Sec. 8.

## 4 Benchmarking and Analysis

Following Table 4, our evaluation largely focuses on the effectiveness of large language models, and to a smaller extent, the performance of our pretrained dual-encoder models at different configurations.

**Dual Encoder Models:** As shown in Table 4, the concat-based fusion technique outperforms the sum-based approach, with BEiT paired with BanglaBERT yielding better performance than the BanglaBERT-ViT combination. Concat-based fusion increased accuracy by 0.18% and 0.42% compared to their sum-based counterparts for BanglaBERT paired with ViT and BEiT encoders, respectively. Using the BEiT encoder increased accuracy by 0.19% compared to ViT for concat-based fusion, but saw a 0.5% decline for sum-based fusion.

**Monolingual vs. Multilingual LLMs:** It can be seen from the benchmarks that models that have an explicit focus on multilingualness, which are models that have been trained on multilingual data, perform better on our dataset compared to monolingual models, largely focused on English. The viability of the tokenizers in a monolingual or multilingual setting also plays an important role [3]. Most of the monolingual models have sub-10% accuracy and LAVE scores, while nearly all the multilingual models make it past that.

**Closed Source LVLMs:** Among the closed source LVLMs, the GPT family outperforms Gemini and Claude with a margin of around 10% in terms of accuracy. Besides accuracy, the performances of Gemini 2.0 Flash, Claude 3.7 Sonnet, GPT-4o, and GPT-4 Turbo are comparable in terms of BERTScore and LAVE, with Gemini having an edge in LAVE. For a rigid metric like accuracy, GPT-4 has a bigger advantage due to the exactness between the answers it generated during QA pair synthesis and the blind answer generation. In open-ended answer generation, metrics like accuracy do not paint the full picture [39].

**Dual Encoder vs. Open Source:** Our dual encoder models outperform all open-source LVLMs, including multilingual ones, across all evaluation metrics. The best-performing model, BanglaBERT-BEiT[Concat], achieves 14.45% accuracy and 20.60% LAVE on the validation set, and 13.10% accuracy and 20.10% LAVE on the test set. These results surpass those of the top-performing zero-shot open-source LVLM, Qwen2.5 VL, which achieves 12.31% accuracy and 17.94% LAVE on the validation set, and 12.04% accuracy and 18.45% LAVE on the test set. While this comparison may not be entirely fair, it suggests that open-source

LVLMs lack sufficient Bangla regional data in their pretraining, indicating the need for fine-tuning to achieve better performance.

**Open Source vs. Closed Source LVLMs:** Based on the benchmarks, the proprietary LLMs surpass the results of open-source models in every evaluation metric. Among the open-source models, Qwen2.5-VL-7B attains the best performance in accuracy and BERTScore. Phi-3.5-Vision, on the other hand, offers the best LAVE score, suggesting that while it does not get the answers as exact as Qwen2.5, it has a more holistic idea of what the correct answer might be, having considered semantic equivalence, synonyms, and variations in verbosity. Proprietary models, such as Gemini 2.0 Flash, Claude 3.7 Sonnet, GPT-4o, and GPT-4 Turbo, are trained on multilingual data. The scale of their training is reflected in the performance improvement over the open-source models.

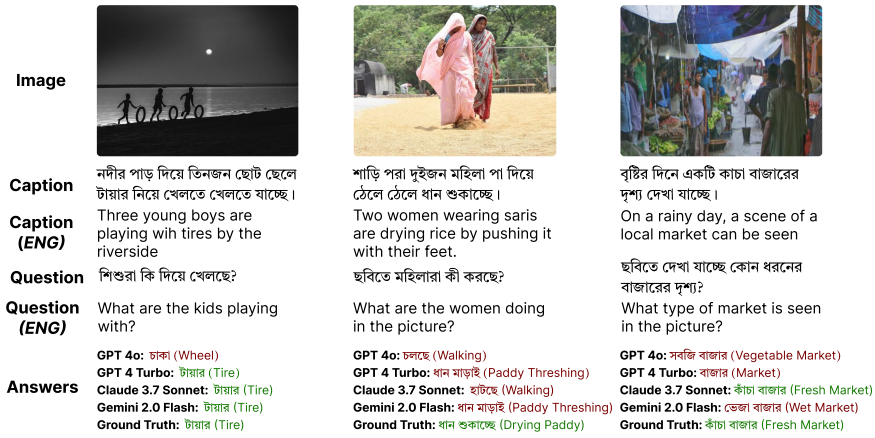
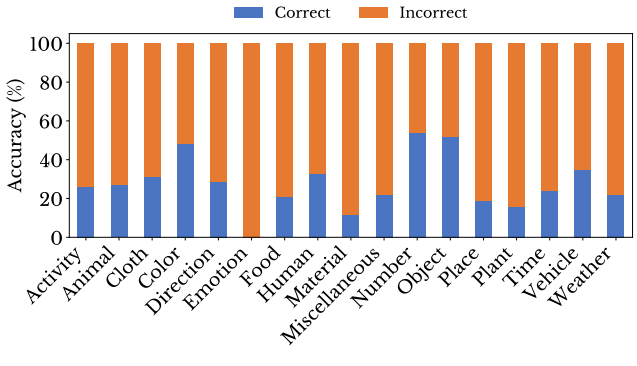



Figure 4: Error Analysis of LLMs in Our Dataset using

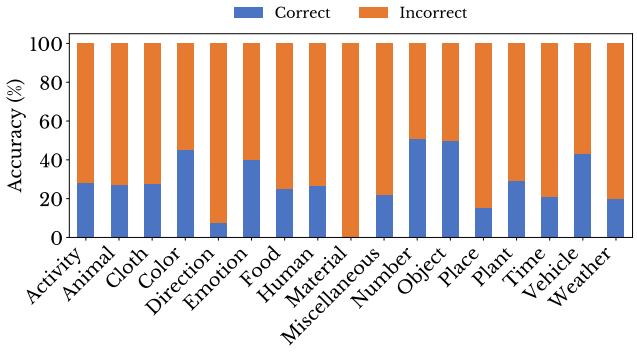
## 5 Error Analysis

From Fig. 4, it can be reasonably inferred that implicit answers can affect the V-LLM performance. In the first image, GPT-4o struggles with the ability to differentiate between two similar-looking objects, failing to tell a "টায়ার" (tire) apart from a "চাকা" (wheel) as both objects are circular. The second example showcases an example of these models lacking regional context in their training data, where the model is unable to recognize the action of drying paddy through manual labor, a common practice in rural Bengal. Finally, the third image shows that using accuracy as a metric treats answers too rigidly, as semantic

equivalence is not assessed. The answers from GPT-4o, GPT-4 Turbo, and Gemini 2.0 Flash are technically acceptable but are considered an error as the ground truth is "কাঁচা বাজার" (fresh market).




(a) Valid

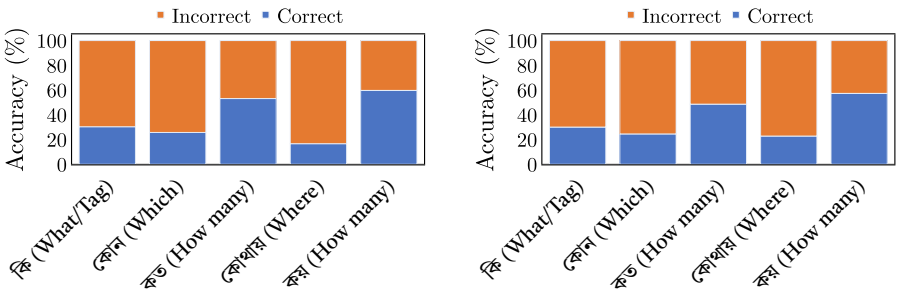



(b) Test

Figure 5: Category-based accuracy by GPT-4 Turbo on valid and test sets

Fig. 5 exhibits that for some of the categories, such as "Color", "Human", "Number", "Object", and "Vehicle", the performance is consistently well across both the validation and test splits. These categories can easily be reasoned with from the images alone, as their presence is explicit. On the other hand, the more abstract or implicit the category is, the harder it gets for the LLM to reason from the visual information alone. This is reflected in categories such as "Emotion", "Direction", and "Time". It should be noted that the performance disparity in "Emotion" is likely due to the low sample count.

For WH-words, GPT-4 Turbo attains 50% or more accuracy on counting questions on both test and validation splits, as seen in Fig. 6. For WH-words




(a) Valid (b) Test

Figure 6: WH-word-based accuracy by GPT-4 Turbo on valid and test sets

such as "কি" (what/tag question) and "কোন" (which), the accuracy is close to the model average. Performance is comparatively poor for "কোথায়" (where), as spatial reasoning from a still image can be challenging. These findings resemble the explicit/implicit subject trend from the categorical breakdown. Note that Fig. 6 only lists the top 5 WH-words (in terms of sample count) for brevity.

## 6 Future Directions

With the growing interest in transliterated text [20], a promising extension of our work is VQA with romanized [16] or code-mixed [2] Bangla. A benchmark incorporating cultural understanding of images, similar to CVQA [45], can also be a potential direction. Finally, our work does not account for regional variations within the Bengal region, which future works can investigate.

## 7 Conclusion

We introduce ChitroJera, a VQA dataset deeply rooted in the geography, culture, and norms of the Bangla-speaking region. To our knowledge, ChitroJera is the first VQA benchmarking with images relevant to the Bengal region, filling a crucial gap in the Bangla vision-language landscape. We anticipate that our work will foster improvements in future models, enabling better modality alignment with low-resource contexts and thereby, overall improved performance.

## 8 Prompts

The prompts used in our literature have been reported here.

**Prompt to Generate the Question-Answer Pairs**

You are an expert in generating Bangla visual question-answer pairs. Given an image and its captions, follow these guidelines:
1. Questions must align with both the image and the captions.
2. Answers should be one or two words, never more than three.
3. Both question and answer must be in Bangla.

⟨CAPTION⟩
Based on the caption and image, generate one pair:
Q: ⟨QUESTION⟩, A: ⟨ANSWER⟩

**Prompt to Generate Answers using GPT-3.5 (Text only model)**

You are an expert Bangla QA assistant. Given a caption and a question, follow these rules:
1. Answer in one or two words (never more than three).
2. Answer must be in Bangla.

⟨CAPTION⟩, ⟨QUESTION⟩
Generate an answer in this format:
A: ⟨ANSWER⟩

**Prompt to Generate Answers with caption**

You are an expert Bangla visual QA assistant. Given an image, its caption, and a question, follow these rules:
1. Answer in one or two words (never more than three).
2. Answer must be in Bangla.

⟨CAPTION⟩, ⟨QUESTION⟩
Generate an answer in this format:
A: ⟨ANSWER⟩

**Prompt to Generate Answers without caption**

You are an expert Bangla visual QA assistant. Given an image and a question, follow these rules:
1. Answer in one or two words (never more than three).
2. Answer must be in Bangla.

⟨QUESTION⟩
Generate an answer in this format:
A: ⟨ANSWER⟩

### $CO_2$ Emission

With a carbon efficiency of 0.432 kg$CO_2$eq/kWh (OECD average), a total of 150 hours of computation was performed using Tesla P100 hardware (TDP of 250W) for the unimodal, multimodal, and dual encoder models. Total emissions of those experiments are estimated to be 16.2 kg$CO_2$eq.

### Acknowledgements

We sincerely appreciate the generous support of our sponsor, Penta Global Limited, Bangladesh, for funding this project.

## References


1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., et al.: Phi-3 technical report: A highly capable language model locally on your phone (2024), https://arxiv.org/abs/2404.14219
2. Alam, S., Ishmam, M.F., Alvee, N.H., Siddique, M.S., Hossain, M.A., Kamal, A.R.M.: Bnsentmix: A diverse bengali-english code-mixed dataset for sentiment analysis. arXiv preprint arXiv:2408.08964 (2024)
3. Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., et al.: Tokenizer choice for LLM training: Negligible or crucial? In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of NAACL 2024. pp. 3907–3924. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: In the IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433 (December 2015)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv (2021)
7. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., et al.: Paligemma: A versatile 3b vlm for transfer (2024)
8. Bhattacharjee, A., Hasan, T., Ahmad, W.U., Samin, K., Islam, M.S., Iqbal, A., Rahman, M.S., Shahriyar, R.: Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. arXiv (2021)
9. Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., Wang, W.: mCLIP: Multilingual CLIP via cross-lingual transfer. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) In the 61st ACL (Jul 2023)
10. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025)
11. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv (2019)
12. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: towards general-purpose vision-language models with instruction tuning. NIPS '23 (2023)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: In NAACL (Jun 2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv (2020)
15. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., et al.: An empirical study of training end-to-end vision-and-language transformers. In: In CVPR. pp. 18166–18176 (2022)
16. Fahim, M., Shifat, F., Haider, F., Barua, D., Sourove, M., Ishmam, M., Bhuiyan, M.: Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 14656–14672 (2024)

17. Faruk, A.M., Faraby, H.A., Azad, M.M., Fedous, M.R., Morol, M.K.: Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In: ICCIT (2020)
18. Gong, Y., Liu, G., Xue, Y., Li, R., Meng, L.: A survey on dataset quality in machine learning. Information and Software Technology **162**, 107268 (2023)
19. Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. 2018 CVPR pp. 3608–3617 (2018)
20. Haider, F., Shifat, F.T., Ishmam, M.F., Barua, D.D., Sourove, M.S.U.R., Fahim, M., Alam, M.F.: Banth: A multi-label hate speech detection dataset for transliterated bangla. arXiv preprint arXiv:2410.13281 (2024)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: In the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: In CVPR (2016)
23. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. arXiv (2021)
24. Hendricks, L.A., Mellor, J., Schneider, Alayrac, J., Nematzadeh, A.: Decoupling the role of data, attention, and losses in multimodal transformers. TACL (2021)
25. Ishmam, M.F., Shovon, M.S.H., Mridha, M., Dey, N.: From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. Information Fusion p. 102270 (2024)
26. Islam, S., Auntor, R., Islam, M., Hossain, M., Islam, A.B.M.A.A., Noor, J.: Note: Towards devising an efficient vqa in the bengali language (06 2022)
27. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (July 2017)
28. Khan, M.F., Shifath, S.S.U.R., Islam, M.S.: BAN-cap: A multi-purpose English-Bangla image descriptions dataset. In: In LREC (Jun 2022)
29. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML. PMLR (2021)
30. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. TMLR (2025)
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
32. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
33. Li, L.H., Yatskar, M., Yin, D.: Visualbert: A simple and performant baseline for vision and language. arXiv (2019)
34. Li, P., Liu, G., He, J., Zhao, Z., Zhong, S.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 374–383. Springer (2023)
35. Libovickỳ, J., Rosa, R., Fraser, A.: How language-neutral is multilingual bert? arXiv preprint arXiv:1911.03310 (2019)
36. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. AI in Medicine (2023)
37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (June 2024)

38. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: In CVPR (2022)
39. Mañas, O., Krojer, B., Agrawal, A.: Improving automatic vqa evaluation using large language models. In: In the AAAI. No. 5 (2024)
40. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: In the CVPR. pp. 3190–3199 (06 2019)
41. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Dataset shift in machine learning (01 2009)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. PMLR
43. Rafi, M., Islam, S., Labib, s.M.H.I., Hasan, S., Shah, F., Ahmed, S.: A deep learning-based bengali visual question answering system (12 2022)
44. Rahman, M., Mohammed, N., Mansoor, N., Momen, S.: Chittron: An automatic bangla image captioning system. Procedia Computer Science (2019), in the ICICT
45. Romero, D., Lyu, C., Wibowo, H.A., others, Aji, A.F.: Cvqa: Culturally-diverse multilingual visual question answering benchmark (2024)
46. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge (2022)
47. Sermanet, P., Ichter, T.D.B., Cao, Y.: Robovqa: Multimodal long-horizon reasoning for robotics (2023)
48. Shah, F., Humaira, M., Jim, M., Ami, A., Paul, S.: Bornon: Bengali image captioning with transformer-based deep learning approach. SN Computer Science **3** (01 2022)
49. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
50. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. PMLR (2019)
51. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Yue, X., Song, Y., Asai, A., Kim, S., et al.: Pangea: A fully open multilingual multimodal LLM for 39 languages. In: ICLR (2025)
54. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
55. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: In CNCCL (Aug 2021)