No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

Youssef Mohamed[1]* **Runjia Li** [2] **Ibrahim Said Ahmad**[3] **Kilichbek Haydarov** [1]
Philip Torr[2] **Kenneth Ward Church** [3] **Mohamed Elhoseiny**[1]*
[1]KAUST [2] University of Oxford [3] Northeastern University

https://www.artelingo.org/

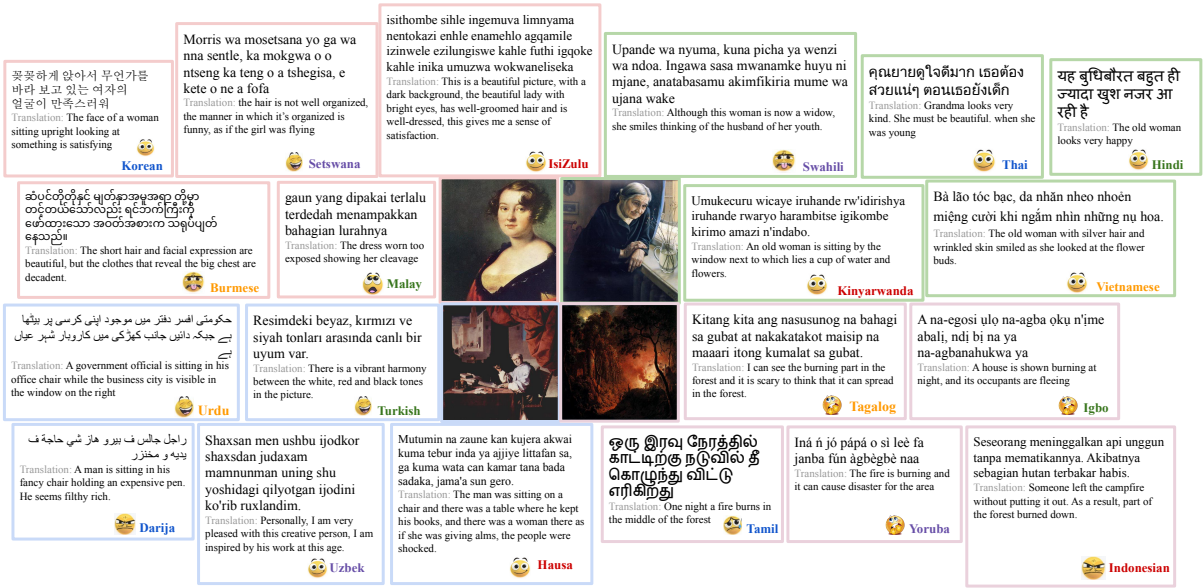

Figure 1: **ArtElingo-28 Benchmark:** 9 emotion labels with captions in 28 languages. The ∼140 annotations per WikiArt image embrace diversity over languages and cultures.

Abstract

Research in vision and language has made considerable progress thanks to benchmarks such as COCO. COCO captions focused on unambiguous facts in English; ArtEmis introduced subjective emotions and ArtELingo introduced some multilinguality (Chinese and Arabic). However we believe there should be more multilinguality. Hence, we present ArtELingo-28, a vision-language benchmark that spans **28** languages and encompasses approximately **200,000** annotations (**140** annotations per image). Traditionally, vision research focused on unambiguous class labels, whereas ArtELingo-28 emphasizes diversity of opinions over languages and cultures. The challenge is to build machine learning systems that assign emotional captions to images. Baseline results will be presented for three novel conditions: Zero-Shot, Few-Shot and One-vs-All Zero-Shot. We find that cross-lingual transfer is more successful for culturally-related languages. Data and code will be made publicly available.

*Corresponding Authors:{fname.lname}@kaust.edu.sa

1 Introduction

A quick review of recent surveys on multimodal AI (Cao et al., 2023; Berrios et al., 2023; Zhang et al., 2023), reveals just how much the literature is focused on English. The literature on benchmarking (Liu et al., 2023c; Li et al., 2023a) provides an astoundingly similar story. With the pervasiveness of AI technology in our societies, it is essential to make the technology accessible to a wider population. Although English is widely spoken as a first language or a second language, most of the world (75% per capita) does not speak English.[1]

Figure 1 shows some annotations from ArtELingo-28. For 2000 images from WikiArt, we have ∼140 emotion labels per image, as well as captions from annotators with diverse backgrounds covering 28 languages. Unlike captions in traditional benchmarks such as COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017) which

[1]https://www.cochrane.org/news/cochrane-evidence-different-languages

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20939–20962
November 12-16, 2024

emphasize unambiguous class labels, the captions in Figure 1 emphasize subjective opinions over objective facts, and diversity over languages and cultures. Figure 1 shows 5 annotations from 5 languages for 4 WikiArt images. Compare, for example, the captions for the first image in Burmese, Malay, Korean and Setswana. There are differences of opinion in both labels and captions:

emotion labels: disgust (Burmese), awe (Malay)
captions: focus on chest (Burmese & Malay);
focus on face and hair (Korean & Setswana)

To advance the field beyond objective facts and unambiguous class labels, it is critical to embrace diversity and subjective differences of opinion. Traditionally, vision research has focused on classifying objects in the image in an objective way, but we prefer to view art as a form of communication between the artist and the audience, where there is more room for subjectivity and diversity. Communication depends on much more than just the pixels in the image such as the cultural backgrounds of the participants.[2]

To add 25 new languages to ArtELingo-28 required considerable effort. Amazon Mechanical Turk works well for a few languages, but less so for many of the 25 languages. ArtELingo-28 consumed 6.25K hours of work, performed by 220 annotators from 23 countries. Compared to ArtELingo which added just 3 languages, our dataset required significantly more management and coordination; ArtELingo-28 was managed by a team of 32 coordinators who contributed more than 2.5K hours.

To cover many practical situations, we utilize ArtELingo-28 to build 3 evaluation setups: Zero-Shot, Few-Shot, and One-vs-All Zero-Shot. The main task evaluates the performance of the generation of affective explanations. In the Zero-Shot setup, we train a model on a large-scale training dataset in a few high-resource languages. We then evaluate that model on languages that do not appear in the the training data. The Few-Shot setup addresses the situation where we have a few training examples in low-resource languages, in addition to the large-scale training dataset from the Zero-Shot setup. We fine-tune the models from the Zero-Shot setup on the few-shot low-resource data and then evaluate them on the rest of the samples. Finally, in the One-vs-All Zero-Shot setup, we have the large-scale training dataset as well as small-scale data in one language (One). After fine-tuning, we evaluate on the Unseen languages (All). This setup is designed to shed light on pairwise interactions between languages, highlighting cultural effects.

We observe clusters (cultural groups) forming from our trained models. These groups go beyond writing systems (scripts), capturing cultural connections between languages.

Additionally, we observe that the multilingual setup is challenging for vision and language models, partly because of the massive vocabulary. We address this challenge by utilizing pretrained multilingual LLMs such as BLOOMZ.

In short, our contributions are:

- We collected 200K emotion labels and affective textual explanations in 25 languages on 2000 images (with ∼140 annotations/image).
- We proposed a benchmark to evaluate the Zero-Shot, Few-Shot, and One-vs-All Zero-Shot performance of Multimodal models.
- We adapt and benchmark four contemporary Vision and Language models to work on our multilingual setup.
- Finally, we study pairwise language transfer revealing insights on cultural differences in emotional perception and expression.

2 Related Work

Multimodal Benchmarks: Benchmarks have always driven the development of many breakthroughs. Imagenet (Deng et al., 2009) being a perfect example, it led to the development of AlexNet (Krizhevsky et al., 2012) which sparked the fire of the deep learning era. Recent benchmarks are moving towards multimodality. In particular, Vision and Language understanding datasets such as COCO (Lin et al., 2014), Conceptual Captions (Sharma et al., 2018), LAION (Schuhmann et al., 2022), VQA (Antol et al., 2015), Visual Commonsense Reasoning (Zellers et al., 2019), and GQA (Hudson and Manning, 2019) have pushed the frontiers of what is possible. They allowed developing models that can perform complex tasks such as visual grounding, image captioning, text-to-image generation, guided segmentation, and more. Although these datasets are framed as benchmarks to develop vision and language, they mainly cover English language.

[2] Blog: Who created the saying that beauty is in the eye of the beholder?

Multilingual Datasets: However, *"English is NOT a synonym for Natural Language Processing,"* Bender (Gradient, 2019). Cultural background has a great influence on perception. The MARVL dataset (Liu et al., 2021) collected concepts and images that are specified by native speakers with diverse backgrounds. Their dataset revealed a major gap in models trained with English-biased datasets such as Imagenet. In addition to MARVL, other multilingual datasets were proposed such as Multi30K (Elliott et al., 2016), a translated version of Flicker30K (Plummer et al., 2015), as well as translated versions of COCO (Rajendran et al., 2015; Yoshikawa et al., 2017; Li et al., 2019; Al-Muzaini et al., 2018). However, translated datasets have much less diversity compared to natively collected data as shown in the MARVL dataset (Liu et al., 2021).

Affective Datasets: The aforementioned datasets describe the facts and objective reality of the visual input. A recent line of work moved beyond factual captions. ArtEmis (Achlioptas et al., 2021; Mohamed et al., 2022b) collected emotion labels and affective explanations to 80K WikiArt artworks. ArtEmis embellished the facts in images with emotions and commentary. The emotional captions revealed new associations that are ignored in factual captions. The subjective captions in ArtEmis exposed the diversity of human responses. ArtELingo (Mohamed et al., 2022a) improved the diversity of affective captioning by including Arabic and Chinese. They showed how the different cultures respond differently to the same visual stimulus. In this work, we embrace the cultural differences by introducing 25 more languages. Table 1 compares three datasets; ArtEmis, ArtELingo, ArtELingo-28.

Apart from the Affective Image Captioning line of work, many other emotions-related datasets were introduced. Unimodal datasets that study emotional responses to single input modality such as text (Strapparava and Mihalcea, 2007; Demszky et al., 2020; Mohammad et al., 2018; Liu et al., 2019), image (Mohammad and Kiritchenko, 2018; Kosti et al., 2017), audio (Cowen et al., 2019, 2020), and multimodal (Busso et al., 2008), however, they are all small scale English datasets. Emotions shape how humans perceive and process external stimuli, and then act upon those signals. Work in the Psychology literature has explored the effect of cultural background on shaping emotional responses (Abu-Lughod, 1990; Henrich et al., 2010; Norenzayan and Heine, 2005). They provide concrete evidence that people from different parts of the world, speaking different languages, perceive the world differently and hence respond differently. ArtELingo-28 is set apart by the inclusion of many languages and hence covering more diverse views of the world.

LLMs: Recently, Large Language Models (LLMs) have become popular due to a number of major successes, driven in large part by the availability of more data and computational power. The power of LLMs became apparent with GPT3 (Brown et al., 2020), a major breakthrough over its predecessors. GPT3 can solve unseen tasks in zero-shot settings. Since then, more and more large language models have been developed; Bloom and OPT (Scao et al., 2022; Zhang et al., 2022) were developed as open-source alternatives to GPT3; Chinchilla (Hoffmann et al., 2022), PALM (Chowdhery et al., 2022), Megatron-Turing NLG (Smith et al., 2022) are proprietary LLMs with even more parameters (more than 175B parameters in GPT). Notably, LLaMa (Touvron et al., 2023) is an open-source model with fewer parameters than GPT3. However, it was trained with more than a trillion tokens. Recently, more powerful models have been developed; Llama2/3 (Touvron et al., 2023), GPT-4 (OpenAI, 2023), and Mistral (Jiang et al., 2023).

Multilingual LLMs: Although the language mentioned above models have achieved impressive results, they mainly focused on English. Multilingual Large Language Models (MLLMs) are a special case of LLMs; unlike LLMs trained in English, MLLMs are pre-trained on text from a more diverse set of languages. MLLMs are more successful than LLMs on tasks involving cross-lingual transfer. For cross-lingual transfer, we fine-tune a pre-trained language model on one language and then apply the model at inference time to unseen languages. This approach offers a number of advantages, especially in low-resource scenarios. XLM-R (Conneau et al., 2019) was one of the first multilingual models to demonstrate cross-lingual transfer capabilities. In the space of Large Language Models, many models make use of pretraining data in a variety of languages. However, English tends to dominate as multilinguality is not the main objective. Bloom (Scao et al., 2022) and mT5 (Xue et al., 2020) are two popular examples of open-source LLM with relatively large contributions of pretraining data from languages beyond English. Building on the success of instruction fine-tuning, Bloomz, and mT0 were introduced (Muennighoff et al., 2022); these are instruction fine-tuned vari-

	ArtEmis	ArtELingo	ArtELingo-28
Image Source	WikiArt	WikiArt	WikiArt
Languages	1	3	3+ **25**
#Images	80k	80k	80k(3) , **2K (25)**
#Annotations	0.45M	1.2M	1.2M(3) + **200K (25)**
#Annot/Image	5.68	15.3	15.3(3) + **140 (25)**
Emotions	9	9	9 (3), **9 (25)**

Table 1: A Comparison of Datasets. ArtELingo-28 extends ArtELingo (Mohamed et al., 2022a) with 200K annotations from 25 additional different languages, many are low-resource.

ants of Bloom and mT5, respectively, with an emphasis on multilinguality. These models exhibit high-quality cross-language transfer. This paper utilizes Bloomz to adapt many Vision and Language models to the multilingual setup.

Vision-LLMs: With LLMs becoming better at generalization, many attempts to integrate modalities were made. For example, in Vision, these models work by injecting visual features into the LLM and then using instruction tuning to teach the LLM reasoning over these features. VisualGPT (Chen et al., 2022) and Flamingo (Alayrac et al., 2022) are two methods that utilized pre-trained LLM and adapted them to visual features produced from a pre-trained vision encoder model. BLIP2 (Li et al., 2023b) proposed using Q-former, a transformer network to map the visual features to the input space of the pre-trained LLM. MiniGPT-4 (Zhu et al., 2023) used BLIP2 architecture with a more powerful language model (Vicuna (Chiang et al., 2023)). They curated a high-quality image-text dataset which resulted in big performance gains over BLIP2. LLaVA (Liu et al., 2023a) used image-text instruction following data to align the output of a frozen image encoder with the input of LLaMa. Finally, InstructBLIP (Dai et al., 2023) created an extensive instruction following the image-text dataset using 26 different open-source datasets. However, None of these methods studied the cross-lingual capability of such methods in a multilingual setting. In parallel to adapting LLMs for visual understanding, some works opted for using smaller models and training them from scratch using losses to properly align vision and language modalities. Notably, X^2-VLM (Zeng et al., 2022a) utilized bounding box descriptions to create better vision and language alignment. CCLM (Zeng et al., 2022b) added a loss function to align the text from multiple languages. Both models achieved very competitive results on the reported benchmarks.

3 ArtELingo-28

Table 1 compares the difference between the three datasets; ArtEmis, ArtELingo, ArtELingo-28. Our dataset ArtELingo-28 expands horizontally by adding more 25 languages. The challenges for such an expansion are unique. We detail in this section our collection effort with a team of 220 native annotators spanning 23 countries.

We utilize the data collection interfaces from ArtELingo (Mohamed et al., 2022a). We ask annotators to carefully examine the image before selecting the most dominant emotion out of 8 emotions.[3] In addition, the annotators have the option to select "Something else" if their feelings do not align with any of the 8 emotions. Finally, the annotators are asked to write an explanation of why the image made them feel the selected emotion. Similar to ArtELingo we *aim* to collect annotations from five different annotators for each image. In total, we cover 2000 images. We make sure to have a representative sample of images covering many genres and styles. (Please see the full list of art styles and genres in appendix C.1.) To embrace the different cultural perspectives, we collect annotations for 25 languages from geographically diverse regions. We cover languages from (we include ArtELingo data for completeness):[4]

- **Africa:** Kinyarwanda, Swahili, IsiZulu, Setswana, Yoruba, Hausa, Igbo, IsiNdebele, IsiXhosa, Emakhuwa.
- **Southeast Asia:** Vietnamese, Indonesian, Thai, Burmese, Malay.
- **Sub-Indian continent:** Tagalog, Tamil, Hindi, Urdu.
- **East Asia:** Korean, *Chinese*.
- **Middle-East:** Turkish, Darija, *Arabic*.
- **Central Asia:** Uzbek, Kazakh, Kyrgyz
- **Europe and North America:** *English*.

Quality Control. We deliver training for our annotators. In their native language, we explain the task and the criteria of good explanations; describing image details and relating those details to the selected emotions. The training includes a clear definition of positive emotion labels since not all cultures agree on the meaning of those labels. We

[3]Positive: Contentment, Awe, Excitement, Amusement
Negative: Sadness, Anger, Fear, Disgust

[4]Please note that some languages are spoken in multiple regions.

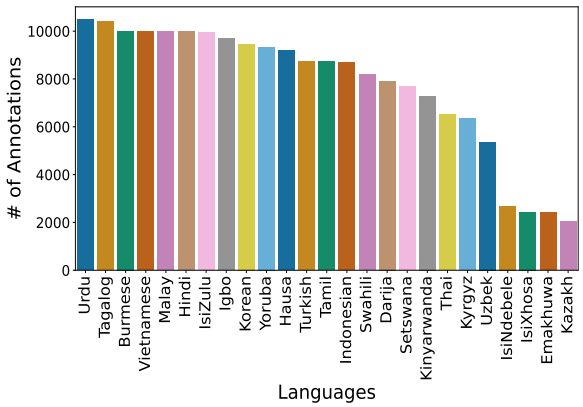

Figure 2: Number of Annotations per Language

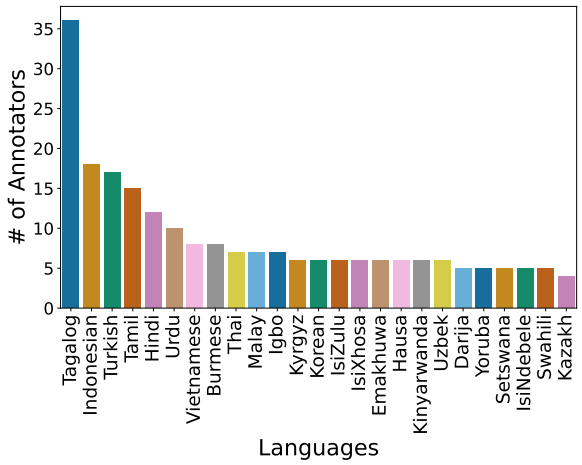

Figure 3: Number of Annotators per Language

provide the detailed descriptions of each emotion in appendix C.3.

In addition, we employ automatic checks that detect duplicates and incorrect language captions. Finally, we hire native speakers as Language Coordinators to perform manual reviewing of the submitted annotations to guarantee high quality. In addition, Language Coordinators are our point of contact with the annotators, they helped us translate the instructions and training content. Appendix B provides more details and statistics about quality control.

3.1 Quantitative Analysis

Number of Annotations. Figure 2 reports the number of annotations per language, ranging from 10,493 for Urdo to 2032 for Kazakh.

Annotators. Figure 3 reports the number of annotators per language. A major challenge for ArtELingo-28 was obtaining access to native speakers, especially for low-resource languages. Amazon Mechanical Turk was used to collect data,

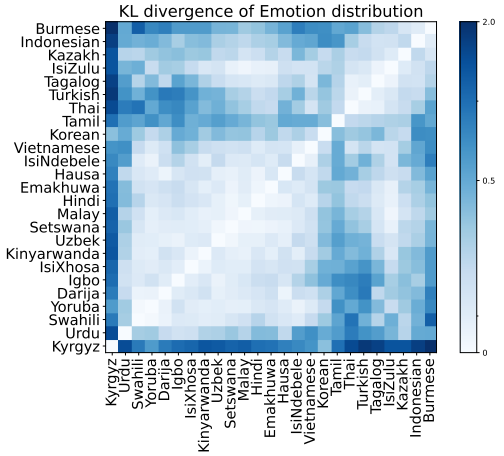

Figure 4: Kullback-Leibler Divergence between the pairwise emotion distribution. The lighter the color the more emotion agreement between the languages.

though not to find annotators. We recruited annotators for Hindi, Urdu, Burmese, Thai, Malay, Vietnamese, Indonesian, Tagalog, Tamil, and Turkish through aiXplain.[5] For other langauges, we used personal networks to find annotators. Although ArtELingo had more annotations, the number of annotators was also proportionally larger making data collection much simpler. For each language, ArtELingo consumed an average of 10.5K hours/language performed by ∼2500 annotators, corresponding to **4.2 hours/annotator**. In ArtELingo-28, we added 25 languages with an average of 250 hours/language and 8.8 annotators corresponding to **28.15 hours/annotator**, seven times as much work per annotator. These calculations do not include management and coordination efforts. In total, annotations consumed **6.25K** hours, plus ∼**2.5K** hours for management hours.

Emotion Distribution. Figure 4 shows KL divergence in emotion labels by language pair. That is, for a pair of languages (l_1, l_2) with emotion distributions (p, q), we calculate the emotional disagreement as $D = \sum_{k \in emotions} p_k * \log \frac{p_k}{q_k}$. We can interpret D as disagreements. The lighter the color, the more similar the language pair. We applied hierarchical clustering to sort languages by agreement in Figure 4. There are two large clusters in the plot; the larger cluster contains mostly languages from Africa and the smaller cluster contains mostly languages from Asia.

[5] https://aixplain.com/

4 Models

We are interested in models able to perform vision and language understanding in many languages. Unfortunately, most open-source state-of-the-art models are heavily biased toward English. Hence, we adapt SOTA general vision and language models to a multilingual setting. The drastic increase in the vocabulary size make such adaptation a challenging task. BLOOMZ tokenizer has a vocabulary size of 250680 tokens compared to only 32000 for Llama2. Hence, the embedding layer of BLOOMZ is much bigger making it harder to align visual features with the language embedding space.

4.1 LLM-based methods

We replace the Large Language Model (LLM) with BLOOMZ and introduce a multilingual instruction-following fine-tuning task, resulting in enhanced performance compared to baseline models. This approach is applied to models such as InstructBLIP (Dai et al., 2023), ClipCap (Mokady et al., 2021), MBlip (Geigle et al., 2023), and MiniGPT-4 (Zhu et al., 2023).

Instructions. We utilize a two-stage training process. The first stage aims to align the visual features with the language model input. In this stage, we utilize large-scale datasets, in particular LAION (Schuhmann et al., 2021) and Conceptual Captions (Sharma et al., 2018), both have English captions only. In addition, we utilize LAION-2B-multi (Schuhmann et al., 2022) which is multilingual. We follow MiniGPT-4 (Zhu et al., 2023) and use the following instructions during the training,

###Human: <Img><ImageHere></Img>Could you describe the contents of this image for me? ###Assistant:

The image embeddings replace the *<ImageHere>* tag. As for LAION-2B-multi, we add "*Use only <language>characters.*" before "*###Assistant:* ". We use the ISO 639-1 standard for naming languages.

In the second stage, we use ArtELingo. To ensure better cross-lingual alignment, we group ArtELingo with image IDs. Then, we randomly sample two languages for the same image. We create instructions in the following format,

###Human: <Img><ImageHere></Img» Could you describe the contents of this image for me? Use <language1>and <language2>words to describe the image. ###Assistant:

We replace *<language1>* and *<language2>* with the two languages we sampled from the dataset. Next, we format the output to be,

<language1>:<cap1>. <language1>:<cap2>

We replace *<cap1>* and *<cap2>* with the captions corresponding to the languages. We found this setup to improve the model's alignment across different languages, and improve generalizations to new languages.

4.2 Non LLM-based methods

Additionally, we adapt models that are not based on LLMs by replacing the language encoder with a multilingual language encoder (XLM-R(Conneau et al., 2019)). CCLM provides pre-trained models with the XLM-R backbone, however, their model does not natively support caption generation since it is an encoder-style model. We follow the standard procedure similar to UniLM (Dong et al., 2019) and X^2-VLM (Zeng et al., 2022a) to generate captions via using the [mask] token autoregressively. In particular, we start with an empty sentence and append [mask] as the first token. The model then predicts a probability distribution over the vocabulary. We sample the first token from that distribution. Next, we append the [mask] token as a second token and repeat the whole process. We continue the generation of tokens until we reach the maximum sentence length or the end of sentence [eos] predicted by the model.

5 Experiments

This section provides baselines for the task of Multilingual Affective Image Captioning. This task takes three inputs: *image, emotion, language*. The model is trained to generate an affective caption in the desired language explaining why the image evokes the desired emotion. The next three subsections discuss three experimental setups to evaluate baseline models over a variety of practical situations. In all three setups, we report standard metrics; BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). We report summary scores averaged over evaluation languages; due to page limitations, detailed results by language are reported in appendix D.3.

Finally, we provide results for emotion label prediction in section 5.4.

Models	BLEU-4	METEOR	ROUGE	CIDEr
InstructBlip	0.54	3.03	5.83	0.05
ClipCap	0.18	0.29	0.23	0.02
mBlip	0.23	0.32	0.73	0.04
MiniGPT-4	**1.09**	**5.8**	**8.47**	**0.33**
CCLM	0.01	0.05	0.02	0.00

Table 2: **Zero-shot Performance.** MiniGPT-4 is the best performing model.

5.1 Setup 1: Zero-Shot

This setup is intended for cases where we have just a few high-resource languages with large training sets. Specifically, we consider Arabic, Chinese, and English as high-resource languages. In the Zero-Shot setting, the system is trained on ArtELingo data (Mohamed et al., 2022a) in Arabic, Chinese and English, and tested on 25 other languages in ArtELingo-28.

Results. Table 2 reports the average scores over all the languages. It is evident that MiniGPT-4 (Zhu et al., 2023) is the best performing model in this setting, followed by InstructBlip. This aligns with results from LVLM-eHub benchmark (Xu et al., 2023) where MiniGPT-4 shows superior performance on open-world scenarios compared to InstructBlip which heavily overfits existing tasks. However, both models are superior to competition due to their superior pretraining strategies.

ClipCap has a similar architecture to MiniGPT-4 and InstructBlip but it does not undergo a similar pretraining. The results show the importance of high-quality pretraining in improving the model's cross-lingual Zero-Shot performance.

Although the CCLM model is reported to achieve SOTA results when trained on a given task (Zeng et al., 2022b), it suffers greatly when it comes to cross-lingual Zero-Shot performance. CCLM is not based on large language models which limits its reasoning, instruction-following, and generalization abilities.

Figure 5 showcases some generations from the MiniGPT-4 model. The quality of the generations is quite good, even on the bottom row where the model was not trained on those on those languages.

5.2 Setup 2: Few-Shot

This setup corresponds to the scenario where we have a modest amount of data (∼7K) from low-resource languages in addition large amounts of data (∼900K) from high-resource languages. We start from the Zero-Shot datasets and the Zero-Shot

Figure 5: **Example Zero-Shot Generations.** The top row is the performance on the test data from ArtELingo where the model has seen the languages during training. The second row corresponds to languages that the model has not seen during multimodal training.

Percentage	BLEU-4	METEOR	ROUGE	CIDEr
20%	**13.5**	14.3	20.4	0.93
60%	13.4	14.2	20.4	0.93
80%	12.9	14.5	20.7	0.95
100%	13.1	**14.6**	**20.9**	0.95

Table 3: **Few-shot Performance.** We observe a significant performance gain on MiniGPT-4 over the Zero-Shot model. MiniGPT-4 is sample efficient and reaches the best performance with few data points due to the reasoning and generalization ability of its LLM.

models, and then further fine-tune the model on the 25 additional languages in ArtELingo-28. We vary the ratio of the training samples from 20% to 100%. Please note that not all 25 languages have the same number of annotations. We report the results from MiniGPT-4 since it is the best performing model.

Results. We report the results in Table 3. We observe a major improvement over the Zero-Shot model's performance. However, we don't observe a significant improvement in the performance as we use more finetuning data. Hence, we recommend collecting more languages over more samples (**expand horizontally**) if the objective is cross-lingual transfer.

5.3 Setup 3: One-vs-All Zero-Shot

This setup aims to study the pairwise interaction between languages. We fine-tune the Zero-Shot model on one language, and evaluate on all the other languages. We hypothesize that successful cross-lingual transfer is driven by close cultural connections. For example, let model x be fine-tuned on Hindi, while model y is fine-tuned on Hausa. We evaluate both models on Urdu. If both models perform **similarly**, then there is **no** underlying cultural relationship between Hindi or Hausa with Urdu. However, if either model performs better, then we can assume an extra cultural connec-

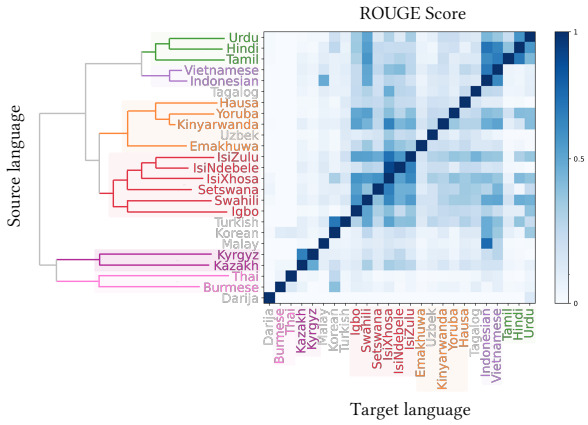

Figure 6: **One vs All Zero-Shot.** The figure shows the rouge score on the target languages. On the left the clustering reveals cultural connections. The captioning scores reveal groups that align with real world cultural connections. This clustering suggests that our trained models can capture the cultural signal.

tion between Urdu and the best performing model's language.

Results. We define the *Source* language as the language used for finetuning the ZeroShot model while the *Target* language is the language being evaluated.

Figure 6 shows ROUGE scores by language pair. Columns are normalized by MinMax scaling.[6] Hierarchical clustering is used to sort languages, so languages with similar scores are grouped near one another. Colors were added manually to call out clusters. By construction, the scores are large along the main diagonal where the same language is used for both fine-tuning and evaluation.

Interestingly, we see language groupings that reflect cultural connections irrespective of other factors such as writing systems. We can see major clusters representing cultural groups. For example:

- Urdu, Hindi, Tamil: These languages share considerable history, though they use quite different scripts. Transfer between these languages is relatively successful compared to other languages.
- Indonesian, Vietnamese: These languages are spoken in Southeast Asia, and are geographically close to one another.
- IsiZulu, IsiNdebele, IsiXhosa, Setswana: These languages are spoken in Southern Africa. They also belong to a larger cluster of African languages: Hausa, Yoruba, Kinyarwanda, Emakhuwa, Swahili, Igbo.
- Kyrgyz, Kazakh: These two languages are similar to each other. Both Kyrgystan and Kazakhstan share considerable history and tradition. Similarly, Burmese and Thai's cultures are related to one another.

We find it interesting to observe such groupings emerging naturally from the data collected from annotators with different backgrounds and traditions. Although some languages share the same writing system characters such as Darija and Urdu, they are very far away in our clustering. This means that the writing system has little to do with this clustering.

We can see the predominant advantage of collecting data from native speakers. Our trained models are better at embracing the different cultural perspectives.

5.4 Emotion Label Prediction

This task takes a caption as an input and predicts one of the nine emotions as an output. We finetune XLM-roBERTa (Conneau et al., 2019) in multiple settings to show the advantages of finetuning with ArtELingo-28. Our first model called *base* is trained on 900K annotations from ArtELingo consisting of captions in Arabic, Chinese, and English languages. *ArtELingo* model is further finetuned on 200K samples from ArtELingo different from the initial training data. This model simulates collecting more data in high resource languages hoping to positively improve multilingual performance. *ArtELingo-28* model load the *base* model and then finetunes on our ArtELingo-28 dataset. This model corresponds to collecting native multilingual data. Finally, $ArtELingo\text{-}28_O$ finetunes XLM-roBERTa only on ArtELingo-28 to measure the usefulness of training using high resource languages before finetuning. In all of our experiments, we finetune the XLM-roBERTa large for 5 epochs.

Table 4 reports the accuracy, precision, recall, and F1 scores for all the models. We evaluate on the test set of ArtELingo-28.

Model	Accuracy	Precision	Recall	F1
base	0.37	0.415	0.325	0.342
ArtELingo	0.354	0.357	0.31	0.313
$ArtELingo\text{-}28_O$	0.636	**0.662**	0.582	0.606
ArtELingo-28	**0.664**	0.651	**0.628**	**0.638**

Table 4: **Emotion Label Prediction.** Finetuning using ArtELingo-28 is essential to learn the culture specific emotional responses.

[6] sklearn.preprocessing.MinMaxScaler

ArtELingo-28 is the best performing model. Notice the huge gap between *ArtELingo* and *ArtELingo-28*. It reflects the need to collect data from native speakers. Naive scaling of data by collecting more samples from the same languages does not help, in contrast, it seems to harm the performance in our case. Finally, the difference between *ArtELingo-28* and $ArtELingo\text{-}28_O$ emphasize the importance of pretraining on a large dataset even if the languages are difference. It shows the ability of the multilingual XLM-roBERTa to do cross-lingual transfer. Appendix D.3.4 include more hyper-parameters and training details.

6 Conclusion

In summary, ArtELingo-28 addresses a critical gap in evaluating large-scale multilingual Affective Vision and Language understanding. By adding 25 languages and 200K high-quality annotations, including low-resource languages, our dataset embraces the cultural differences.

Our evaluation setups — Zero-Shot, Few-Shot, and One vs All Zero-Shot — assess affective explanation generation across diverse linguistic contexts. The One vs All Zero-Shot setup extends evaluation to languages beyond the training dataset, revealing cultural connections through cross-lingual transfer performance.

In this work, we introduced a dataset, proposed a benchmark, and adapted four Vision and Language models, overcoming current limitations in multilingual AI evaluation. ArtELingo-28 sets a benchmark for bridging linguistic and cultural gaps in Affective Vision and Language understanding.

7 Limitations

Modeling. Our approach faces limitations in modeling, evaluation, and data. Modeling relies on the quality and availability of multilingual large language models (MLLMs), stressing the need for attention and resources beyond English models. Current evaluation metrics overlook emotional and subjective aspects, necessitating new metrics. To enhance benchmarking, broader language coverage in evaluation datasets is crucial, alongside the collection of more diverse native multilingual data. Addressing these limitations is essential for advancing Multilingual General Purpose Vision Language Models' effectiveness and applicability.

Dataset. In addition to the limitations of ArtELingo (Mohamed et al., 2022a), our dataset reflects the viewpoints of the annotators. We instructed the annotators neither to attack any given group of people based on ethnicity, religion, etc. nor use vulgar language. In addition, We asked our coordinators to reject and report any captions that used vulgar language or attacked a group of people.

The captions might reflect ideas that might be sensitive in the western cultures. However, they reflect the world views of people from different cultures. We should expect some things and topics to be more or less sensitive depending on culture: dress, gender, religion, drinking, sex, respect for animals, respect for the environment, etc. It is not helpful to demand that all cultures conform to a specific world view.

Our annotators represent a group of people who have internet access and are educated. Most of them speak English in addition to their native language. We leave extending our dataset to include other groups of people to future work.

Our goal in ArtELingo-28 is to embrace diversity and capture the diverse perceptions of the world held by different cultures. While we may disagree on topics such as religion, dress, and other cultural norms, our position as authors, coming from different cultures and covering three continents, is that it is important to consider these varying perspectives to build culture-aware models, while of-course promoting respect and eliminating, as much as we can, hateful content. We should embrace the fact that what is considered appropriate or inappropriate varies across cultures and try to understand other people. We find it not our job to draw one line for all cultures but to expose this phenomenon that we find worth studying. Our dataset aims to help people understand and respect each other's worldviews, even if disagreement is inevitable on some topics, which may serve as a resource for advancing cultural and cross-cultural psychology.

8 Acknowledgment

The authors would like to thank aiXplain for annotators in 10 languages and providing high quality annotations. In addition, we thank all our language coordinators for their amazing effort and support throughout the project. We extend our gratitude to all the annotators for their effort in the data collection.

This work was supported by King Abdullah University of Science and Technology (KAUST), under Award No. URF/1/5016.

References

Lila Abu-Lughod. 1990. The romance of resistance: Tracing transformations of power through bedouin women. *American ethnologist*, 17(1):41–55.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

Huda A Al-Muzaini, Tasniem N Al-Yahya, and Hafida Benhidour. 2018. Automatic arabic image captioning using rnn-lstm-based language model and cnn. *International Journal of Advanced Computer Science and Applications*, 9(6).

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Alan S Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2019. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698.

Alan S Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. 2020. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117(4):1924–1934.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

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

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

The Gradient. 2019. The bender rule: On naming the languages we study and why it matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/. Accessed: November 17, 2023.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2017. EMOTIC: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298, Hong Kong, China. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022a. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. *arXiv preprint arXiv:2211.10780*.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. 2022b. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21272.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. WikiArt emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ara Norenzayan and Steven J Heine. 2005. Psychological universals: What are they and how can we know? *Psychological bulletin*, 131(5):763.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2015. Bridge correlational neural networks for multilingual multimodal representation learning. *CoRR*, abs/1510.03519.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022a. X^2-VLM: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2022b. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv preprint arXiv:2206.00621*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xinsong Zhang, Yan Zeng, Jipeng Zhang, and Hang Li. 2023. Toward building general foundation models for language, vision, and vision-language understanding tasks. *arXiv preprint arXiv:2301.05065*.

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

A Dataset Collection Interface

We use the interface in Figure 7 to collect the annotations for ArtELingo-23. We hire native speakers who are very proficient in English to translate the interface to the respective languages.

B Quality Control

Our quality control includes multiple stages. Below we outline these stages and their details:

- **Hiring Stage:** we have worked with our language coordinators on multiple projects before and they have a proven track record of providing high quality data through their annotators. They know the annotators in person as a result the hired annotators are top quality.
- **Training Stage:** we provide detailed training through a conference meeting for all the annotators and coordinators. We then ask the coordinators to translate the list of instructions and the key points of the training into their respective language to make sure that all annotators understand the instructions completely.
- **Reviewing Stage:**
  - We developed automatic scripts that perform duplicate detection, sentiment analysis verification, and language identification. For duplicates we used exact text matching. We utilized a sentiment analysis model fine tuned on ArtELingo (English, Arabic, Chinese) to classify whether the caption is positive or negative. Finally, we utilized the NLLB(Costa-jussà et al., 2022) language identification fasttext model [7] to make sure the languages match our target. Any automatically rejected annotation was marked for the language coordinators to review in detail. Noteworthy, duplicate detection was triggered 20%, 32%, 47% for Thai, Urdu, and Turkish (All languages were supported by AiXplain). We discussed the issue with the annotators; All duplicates were re-annotated and we didn't observe this issue. The sentiment analysis classifier marked very few instances <1%. Finally, the language identification also marked <1% for all languages except Malay and Indonesian since the two languages are very similar to each other. The language identification model classified 75% of Malay instances as indonesian. However, manual inspection revealed no issues.
  - The language coordinators manually review the annotations. Overall, the rejection rate in each language was between 1 5% reflecting the high quality of annotations. The most common mistake encountered was selecting the "something else" emotion label while the caption reflects one of the 8 emotions. For this mistake, we provided extra training that focused on explaining the emotional labels for the annotators. The initial training explained the emotion labels but some languages (Burmese, Turkish, Swahili, Hausa, Indonesian, Korean) required extra training. After that the annotators re-selected the emotional label to align better with our definition of emotions.
  - Another issue is that some annotators started their sentences with "This image makes me feel ..." which heavily influenced the performance of our captioning models. We asked the annotators to fix those annotations by paraphrasing them to more natural sentences. This was mainly encountered in Yoruba, but was fixed early in the annotation process.
- **Translated Validation:** As a final quality check after the coordinators, we (authors) translated a random sample of 500 annotations from each language to English and performed sanity checks. We didn't encounter any bad quality annotations.

C Dataset Analysis

C.1 Art Styles and Genres

In total, we cover a subset of 2000 images from the set of 80K images used in ArtELingo. We make sure to have a representative sample of images covering all the art styles and genres.

[7] huggingface.co/facebook/fasttext-language-identification

Emotion	Description
Contentment	"A deep sense of satisfaction and inner peace. It often arises when a person feels comfortable, secure, and at ease with their present circumstances. *Example Caption: A peaceful, sunlit afternoon by the lake, with a person sitting on a comfortable chair, sipping tea, and smiling at the serene view of nature.*"
Excitement	"An intense feeling of enthusiasm, eagerness, and heightened energy. It typically emerges in response to something thrilling or anticipated, such as a special event, achievement, or adventure. It often involves a desire to engage fully in the exciting experience. *Example Caption: A joyful crowd at a music festival, hands raised, faces beaming with exhilaration, as colorful confetti rains down on them during a thrilling performance.*"
Amusement	"A light-hearted emotion associated with joy and laughter. It arises when something is funny, entertaining, or amusing. It often involves a response to humor, jokes, or playful situations. *Example Caption: Friends gathered around a table, laughing uncontrollably as they played a hilarious board game, with one person wearing a funny costume and others doubled over in laughter.*"
Awe	"Experienced when encountering something vast, magnificent, or transcendent. It involves a sense of wonder, reverence, and humility in the face of something that is awe-inspiring. Awe can be triggered by natural phenomena like a breathtaking landscape, the night sky, or by human achievements that evoke a sense of grandeur and beauty. *Example Caption: A breathtaking sunset over a majestic mountain range, casting vibrant hues of orange and purple across the sky, leaving a lone observer standing in awe of nature's grandeur.*"

Table 5: Description of emotion labels

The art styles are Abstract Expressionism, Action painting, Analytical Cubism, Art Nouveau Modern, Baroque, Color Field Painting, Contemporary Realism, Cubism, Early Renaissance, Expressionism, Fauvism, High Renaissance, Impressionism, Mannerism Late Renaissance, Minimalism, Naive Art Primitivism, New Realism, Northern Renaissance, Pointillism, Pop Art, Post Impressionism, Realism, Rococo, Romanticism, Symbolism, Synthetic Cubism, and Ukiyo-e.

While the genres are portrait, landscape, genre painting (misc), religious painting, abstract painting, cityscape, sketch and study, still life, and illustration.

C.2 Caption Length

Figure 8 showed the number of characters and number of bytes per caption in separate plots. We combine both measures in figure 9 to better understand the effect of character encoding. Most languages have approximately one to one byte to character ratios. While other languages such as Korean and Thai have higher ratios, approximately four and three, respectively. It is interesting to study the effect of this ratio on the reported metrics in the caption generation experiments. Although it is difficult to disentangle the effect of different confounders such as the LLM pretraining, tokenizer, etc.

C.3 Emotion Distribution

Figure 10 shows the emotion distribution for the different languages. Although we provided a detailed description of each emotion label (see Table 5), we still see variations in some languages.

C.4 Qualitative Samples

We show different annotations from ArtELingo-23 in Figure 11. More examples can be found in our repository.

C.5 Ethical Concerns

We received approval for the data collection from KAUST Institutional Review Board (IRB). The IRB requires informed consent; in addition, there are terms of service in AMT. We respected fair treatment concerns from EMNLP (compensation) and IRB (privacy).

The workers were given full-text instructions on how to complete tasks, including examples of approved and rejected annotations. Participants' approvals were obtained ahead of participation. Due to privacy concerns from IRB, comprehensive demographic information could not be obtained.

We compensated all annotators with $0.1 per annotation making the total $200. The time taken per annotation is on average 45 seconds making the hourly payment $8 / hour which is above the minimum wage in all of the respective countries. For the language coordinators, we compensated them with $200 for their reviewing and communication efforts. The workload was lower for the coordinators per annotation. For the head coordinators (hiring and managing multiple language coordinators), they were included as co-authors in this paper.

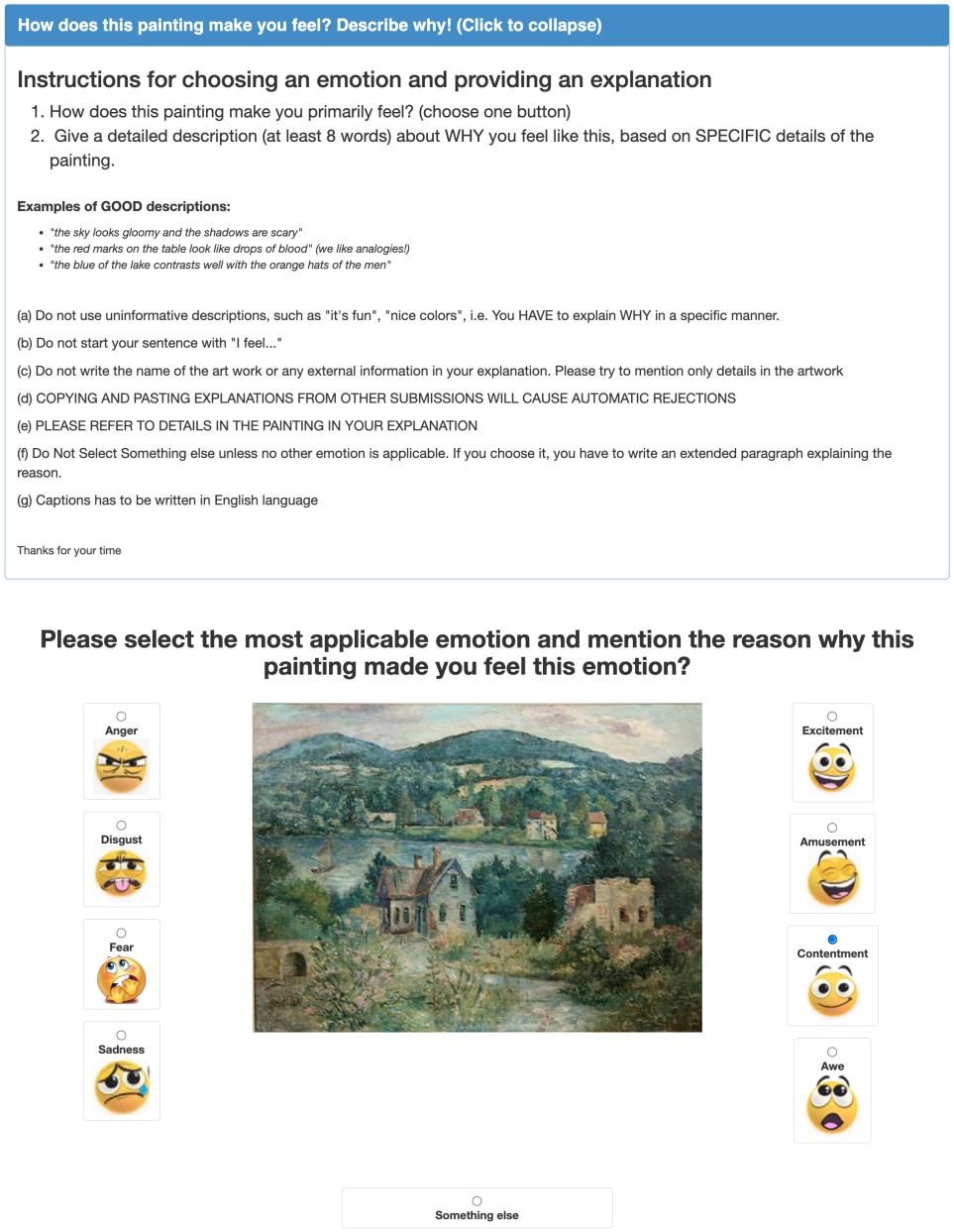

Figure 7: Interface used to collect ArtELingo-23

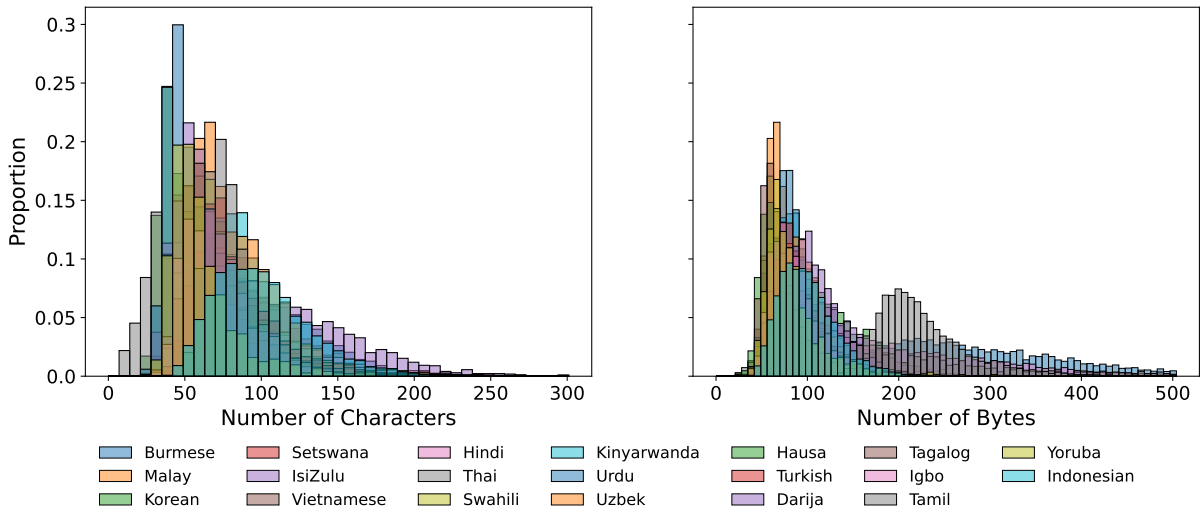

Figure 8: **Left:** Distribution of the caption length as the number of characters. **Right:** Distribution of the caption length as number of bytes required for storage.

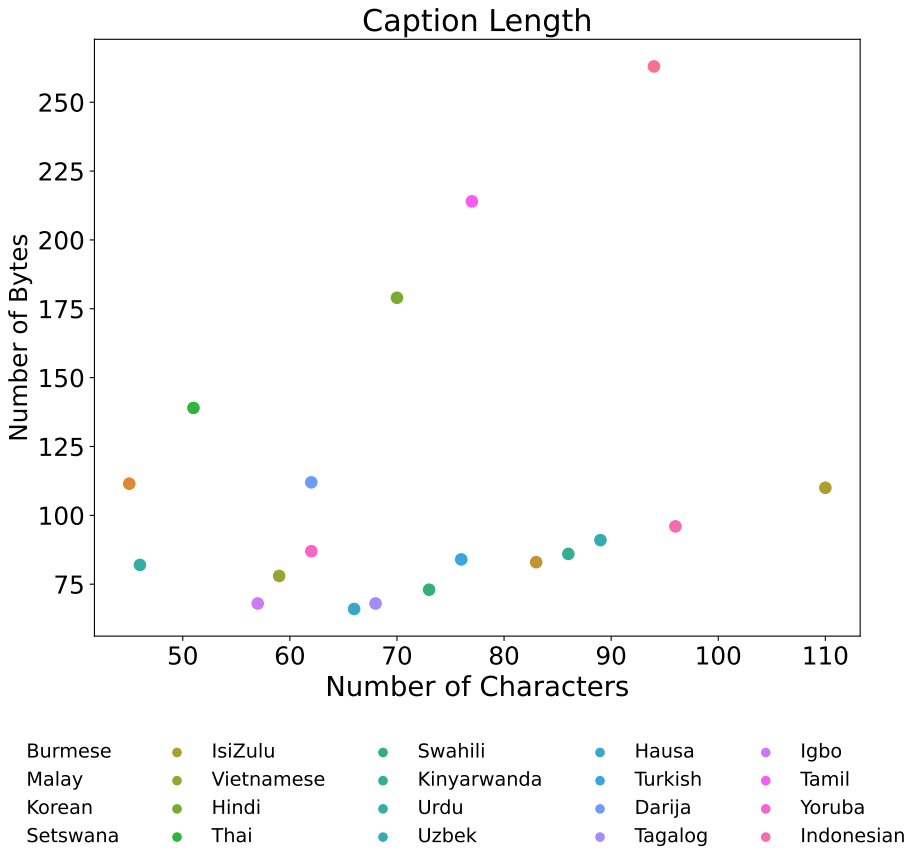

Figure 9: Median Number of Characters vs Number of Bytes per caption

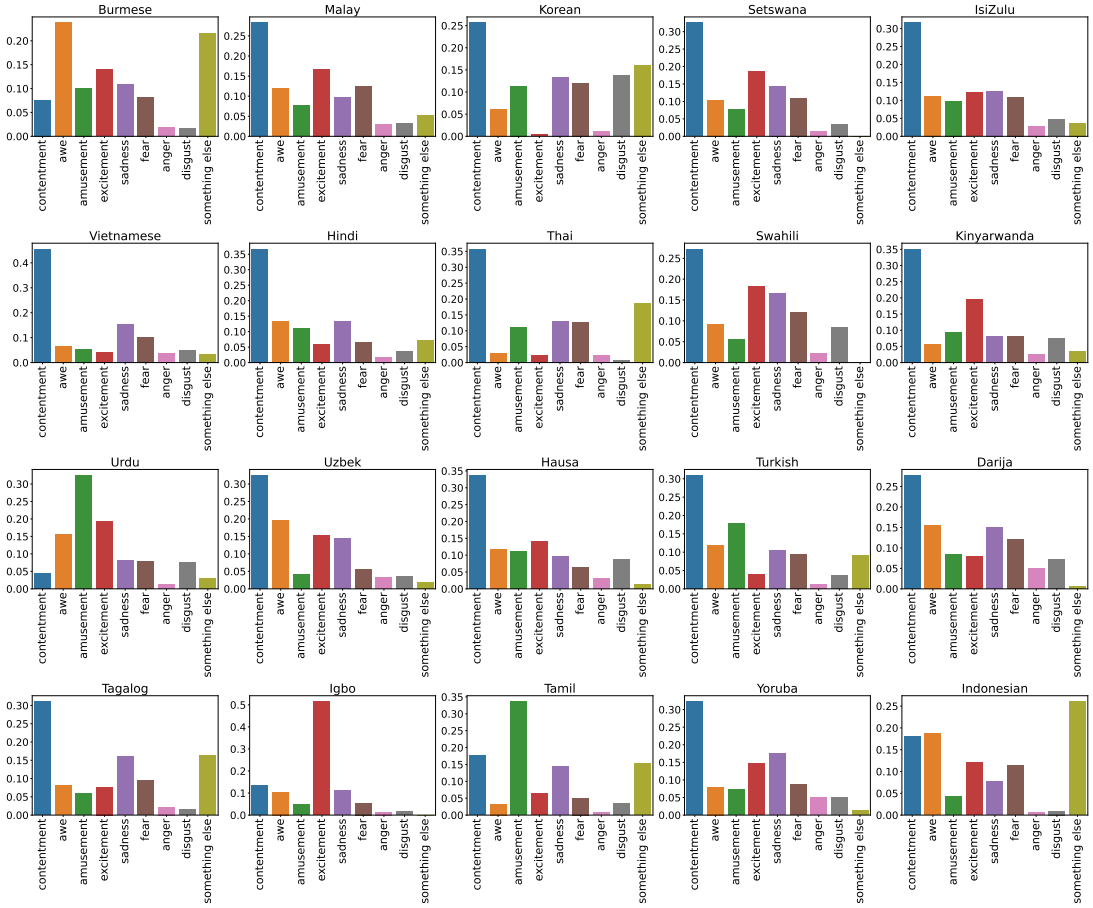

Figure 10: Emotion Distribution for different Languages

العنق الطويل ديال الرجل تيبان بحال شي زرافة تيضحك **Darija**	လူ တစ် ယောက် ၏ အသွင် အပြင် အား အမြင် အာရုံ အတိုင်း ရေး ဆွဲ ထား ခြင်း မျိုး မဟုတ် ပဲ ခံစား ချက် အလိုက် ရေး ဆွဲ ထား ခြင်း မျိုး ဖြစ်သည်။ **Burmese**	Bai min kyau ba sabida yadda aka zana masa dogon wuya yayi yawa **Hausa**
एक आदमी अधिकारी कपड़े पहने हुए, कुर्सी पर संतोष से बैठा है और सामने देख रहा है। उसके चेहरे पर संतोष का भाव प्रकट होता है। **Hindi**	Anya nwaokoroọbịa a dịka nke onye dị nwaayọ, ma ọnụ ya dịka nke onye na-achọ okwu **Igbo**	Leher pria pada gambar sangat panjang. Tidak seperti leher manusia pada umumnya. **Indonesian**
Indoda esesithombeni ingenza ngiphatheke kabi, ngoba ubukeka ecebile futhi ezicabangela yena yedwa, akanandaba nalutho ngaphandle kwakhe. **IsiZulu**	Umugabo wambaye ikoti ry'umutuku yicaye imbere y'urukuta rusize irangi ry'umweru. **Kinyarwanda**	유난히 긴 목과 비대칭을 이루고 있는 얼굴 의 모습이 보기 좋지 않습니다 **Korean**
Seorang lelaki memakai baju kot merah sedang duduk bersendirian **Malay**	O apere sentle monna yo, e bile o bonala jaaka motho yo a tshelang sentle ka kagiso. **Setswana**	mwanamume ameketi kwa upole kwenye kiti,nyuma yake kuna kitambaa cha rangi ya buluu ambayo inaonekana vuzuri na suti nyekundu **Swahili**
ang leeg ng lalaki ay mahaba at ito ay katawa tawa **Tagalog**	அழகான ஆடை அணிந்த ஒருவள் நாற்காலியில் சோகத்துடன் அமர்ந்துக் கொண்டு இருக்கிறாள் **Tamil**	ดวงตาและรูปร่างของผู้ชายดูน่ากลัว **Thai**
Adam sakin görünüyor. Kıyafetinin kırmızı, siyah ve beyaz tonları arasında güzel bir uyum var. **Turkish**	نیلی آنکھیں پرسکون چہرا والا مرد خوشی کا احساس دیلاتا ہے **Urdu**	Hotirjam o’tirgan erkak o’ta mahorat bilan ishlangan. Umumiy portret ko’rinishi quvoch beradi. **Uzbek**

Figure 11: An annotation from ArtELingo-23. Notice the diverse expressions and different points of view across languages.

D Evaluation

D.1 CLIPScore

We attempted to use CLIPscore (Hessel et al., 2021) to evaluate the generations from our trained model. Since CLIPScore is defined for English language only, we used NLLB (Costa-jussà et al., 2022) 1.3B model to translate all of our generations into English then evaluated them using CLIPScore. Table 6 reports the scores for InstructBLIP, MiniGPT-4, and ClipCap. According to the results ClipCap is the best performing model. However, upon qualitative inspection of the model output, we observed the very low quality of CLIPCap captions where the generated text is a repetition of the same word. We also observed that the translation model introduce new words that are not faithful to the original translation which might explain the difference in scores (Such analysis was performed on Arabic language only). Accordingly, we decided not to include the results in the main paper.

Model	CLIPScore	RefCLIPScore
InstructBlip	0.5879	0.6566
ClipCap	**0.6111**	**0.6578**
MiniGPT-4	0.4791	0.5648

Table 6: **CLIPScores and RefCLIPScores.** Zero-Shot performance evaluate using CLIPScore.

D.2 GPT Score

We utilized GPT-4o-mini to evaluate the quality of the generated captions. GPT4 has been shown to align with human judgement (Liu et al., 2023b). Due to cost and rate limitations associated with GPT4 API, we sampled 100 captions from each model in the Zero-Shot performance. We provided the following prompt to GPT-4o-mini:

```
You will give a score between 1 (worst) and 5 (best) for the
   following caption and image. Your score will reflect how much
   the caption is faithful to the image. If the caption is not in
    {lang} language, give a score of 0. Output only the score
   without any explanation. The caption is: {caption}
```

Table 7 reports the GPTScores. Since the conclusion is similar to the traditional metrics and we only used 100 samples from each model, we decided not to include the results in the main paper to avoid confusion related to sample size.

Model	GPTScore
InstructBlip	0.071
ClipCap	0.61
MiniGPT-4	**0.79**

Table 7: **GPTScores.** Zero-Shot performance evaluate using GPTScore.

D.3 Per Language Evaluation

In this section, we dive deeper into the results reported in the experiments section. We break down the result of each experimental setup by language. We don't report the complete results of CCLM since it didn't generalize to any unseen languages. This shows the importance of Multilingual Large Language Models in creating universal multimodal models that perform well on unseen languages.

We utilize these results to better understand the limitations of current models. We observe a major bottleneck in the performance of the underlying Multilingual Large Language Model. This is attributed to the lack of powerful counterparts to the English LLMs. We believe that Multilingual LLMs are

essential for multilingual multimodal generalization. Hence, we encourage the community to develop more powerful Multilingual LLMs compared to Bloomz.

We face a problem where most metrics' implementation is designed for the English language only. While some implementations can work in other languages where words are separated by white space, this is not generally the case. Especially, for languages like Burmese where the separation between semantic tokens is not as straightforward. As a workaround, we utilize the multilingual tokenizer from the Bloomz model to divide the sentence into tokens. Then, we re-join the produced tokens with whitespaces. We found that these methods work well in most languages.

Nonetheless, such a workaround is not ideal since the tokenizers are usually biased towards languages written in Latin script giving them larger semantic tokens. On the other hand, we found that languages such as Burmese and Vietnamese are tokenized into almost character level. A direct consequence is that the scores scale is drastically different making it impractical to compare performance across different languages. We believe there is a huge room for improvement in multilingual tokenization. A very useful feature would be to guarantee a similar token length across different languages.

In the Few-shot and Seen-Unseen experiments, we used MiniGPt4 since it has the fastest training time as well the best generalization performance.

D.3.1 Zero-shot Setting

Tables 8, 9, 10, 11 report the BLEU, ROUGE, METEOR, and CIDEr scores, respectively. We can immediately notice the poor performance on Burmese and Thai. This reflects the inability of the Bloomz LLM to speak in those languages without finetuning. We can also see that the performance of different models varies greatly in some languages. MiniGPT4 performs the best overall closely followed by clipcap and then InstructBlip.

Model	Burmese	Malay	Korean	Setswana	IsiZulu	Vietnamese	Hindi	Thai	Swahili
ClipCap	0.002	0.506	0.18	0.877	0.39	6.544	4.107	0.001	1.109
InstructBlip	0.001	0.449	0.028	0.572	0.356	0.673	0.097	0.0	0.58
MiniGPT4	0.002	0.337	0.979	1.223	0.635	4.21	4.344	0.0	2.097
Model	Kinyarwanda	Urdu	Uzbek	Hausa	Kazakh	Turkish	Darija	Tagalog	Kyrgyz
ClipCap	0.586	2.382	0.271	0.702	0.292	0.498	0.762	0.739	0.299
InstructBlip	0.401	0.351	0.257	0.713	0.315	0.663	0.002	0.737	0.407
MiniGPT4	1.26	2.474	0.166	0.424	0.151	0.164	0.007	0.594	0.163
Model	IsiNdebele	Igbo	Emakhuwa	IsiXhosa	Tamil	Yoruba	Indonesian		
ClipCap	0.518	0.393	0.295	0.399	1.93	0.157	4.713		
InstructBlip	0.327	0.752	0.368	0.342	0.106	0.28	0.615		
MiniGPT4	0.381	1.829	0.352	0.358	2.141	0.62	2.943		

Table 8: Zero-shot Performance on BLEU-4

Model	Burmese	Malay	Korean	Setswana	IsiZulu	Vietnamese	Hindi	Thai	Swahili
ClipCap	0.014	3.653	1.202	5.221	2.376	20.723	14.437	0.005	5.723
InstructBlip	0.006	1.78	0.154	3.04	2.07	2.257	0.329	0.0	2.801
MiniGPT4	0.017	1.312	2.809	5.3	2.765	12.96	13.978	0.0	7.225
Model	Kinyarwanda	Urdu	Uzbek	Hausa	Kazakh	Turkish	Darija	Tagalog	Kyrgyz
ClipCap	3.197	9.421	1.767	3.845	1.994	3.111	3.166	4.044	1.984
InstructBlip	2.042	1.317	1.444	3.267	1.915	3.327	0.012	3.433	1.939
MiniGPT4	5.083	9.076	0.931	1.971	0.925	0.819	0.032	2.679	0.926
Model	IsiNdebele	Igbo	Emakhuwa	IsiXhosa	Tamil	Yoruba	Indonesian		
ClipCap	3.396	2.668	1.739	2.316	7.648	0.741	17.098		
InstructBlip	1.866	3.624	1.746	1.819	0.422	1.234	2.661		
MiniGPT4	1.957	6.584	1.594	1.825	7.257	2.307	10.854		

Table 9: Zero-shot Performance on ROUGE

Model	Burmese	Malay	Korean	Setswana	IsiZulu	Vietnamese	Hindi	Thai	Swahili
ClipCap	0.054	2.122	0.989	2.208	1.111	23.151	10.699	0.316	3.307
InstructBlip	0.003	0.934	0.452	1.352	0.986	5.336	0.126	0.151	1.367
MiniGPT4	0.047	0.712	2.53	3.235	1.617	18.069	10.929	0.216	4.911
Model	Kinyarwanda	Urdu	Uzbek	Hausa	Kazakh	Turkish	Darija	Tagalog	Kyrgyz
ClipCap	1.439	7.312	0.854	1.707	0.827	1.36	2.029	1.798	0.917
InstructBlip	0.888	0.755	0.713	1.603	0.65	1.52	0.005	1.748	0.901
MiniGPT4	3.922	6.976	0.496	1.039	0.319	0.4	0.016	1.469	0.357
Model	IsiNdebele	Igbo	Emakhuwa	IsiXhosa	Tamil	Yoruba	Indonesian		
ClipCap	1.456	2.063	0.831	1.025	5.239	0.415	10.429		
InstructBlip	0.902	1.28	0.888	0.833	0.197	0.552	1.275		
MiniGPT4	0.984	8.778	0.795	0.836	5.294	2.344	6.544		

Table 10: Zero-shot Performance on METEOR

Model	Burmese	Malay	Korean	Setswana	IsiZulu	Vietnamese	Hindi	Thai	Swahili
ClipCap	0.0	0.276	0.001	0.084	0.026	17.113	5.874	0.0	1.049
InstructBlip	0.0	0.09	0.0	0.075	0.061	2.169	0.004	0.0	0.087
MiniGPT4	0.0	0.157	0.372	0.984	0.523	13.53	5.568	0.0	3.311
Model	Kinyarwanda	Urdu	Uzbek	Hausa	Kazakh	Turkish	Darija	Tagalog	Kyrgyz
ClipCap	0.029	4.989	0.018	0.038	0.0	0.022	0.956	0.108	0.0
InstructBlip	0.071	0.746	0.032	0.188	0.0	0.115	0.003	0.34	0.01
MiniGPT4	2.219	5.628	0.026	0.209	0.0	0.035	0.012	0.52	0.002
Model	IsiNdebele	Igbo	Emakhuwa	IsiXhosa	Tamil	Yoruba	Indonesian		
ClipCap	0.007	0.138	0.098	0.054	6.424	0.12	8.712		
InstructBlip	0.045	0.088	0.14	0.049	0.043	0.205	0.246		
MiniGPT4	0.209	1.206	0.202	0.147	6.251	0.769	5.955		

Table 11: Zero-shot Performance on CIDEr

D.3.2 Few-shot Setting

We report the extended results for the One-vs-All Zero-Shot setting in this anonymous file: https://github.com/Mo-youssef/artelingo-28/tree/main/results/minigpt/fewshot.csv

D.3.3 One-vs-All Zero-Shot Setting

We report the extended results for the One-vs-All Zero-Shot setting in this anonymous file: https://github.com/Mo-youssef/artelingo-28/tree/main/results/minigpt/seenunseen.csv

D.3.4 Emotion Label Prediction

We utilize XLM-roBERTa large model from huggingface[8]. We finetune the model for 5 epochs in all of our experiments. We use AdamW optimizer with $LR = 2e - 5$ and $eps = 1e - 8$ with a linear decay scheduler. We used a batch size of 128 and max caption length of 128 tokens where padding and truncation are utilized to fix the batches number of tokens. The gradients were clipped to have a norm of 1.

[8]https://huggingface.co/FacebookAI/xlm-roberta-large