CROPE: Evaluating In-Context Adaptation of Vision and Language Models to Culture-Specific Concepts

Malvina Nikandrou Georgios Pantazopoulos Nikolas Vitsakis

Ioannis Konstas Alessandro Suglia

Heriot-Watt University
{mn2002, gmp2000, nv2006, i.konstas, a.suglia}@hw.ac.uk

Abstract

As Vision and Language models (VLMs) are reaching users across the globe, assessing their cultural understanding has become a critical challenge. In this paper, we introduce CROPE, a visual question answering benchmark designed to probe the knowledge of culture-specific concepts and evaluate the capacity for cultural adaptation through contextual information. This allows us to distinguish between parametric knowledge acquired during training and contextual knowledge provided during inference via visual and textual descriptions. Our evaluation of several state-of-the-art open VLMs shows large performance disparities between culture-specific and common concepts in the parametric setting. Moreover, experiments with contextual knowledge indicate that models struggle to effectively utilize multimodal information and bind culture-specific concepts to their depictions. Our findings reveal limitations in the cultural understanding and adaptability of current VLMs that need to be addressed toward more culturally inclusive models.[1]

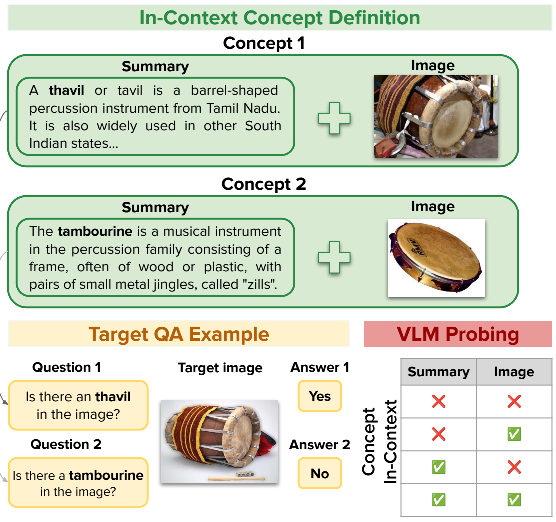

Figure 1: CROPE probes the cultural knowledge of VLMs and assesses the effect of contextual information. Each dataset sample poses a question about the presence of a culture-specific concept within an image and is paired with demonstrative text and images that can be used as additional context to improve understanding.

1 Introduction

Recent Vision and Language models (VLMs) (Wang et al., 2024a; Laurençon et al., 2024; Li et al., 2024a) have shown impressive performance across a variety of benchmarks (Li et al., 2023; Yu et al., 2024). At the same time, frontier VLMs (Achiam et al., 2023; Team et al., 2023) have become widely accessible, making it crucial that these models can grasp the nuances of different cultures. Models lacking cultural awareness can affect global cultural diversity, as they can potentially contribute to content reinforcing beliefs, habits, or perspectives from more dominant cultures (Arora et al., 2023; Cao et al., 2023; Tao et al., 2023).

Cultural concepts encompass both universal categories, such as birthdays, weddings, and funerals (Acharya et al., 2020), as well as specific concepts that are primarily encountered within a particular community. In this work, we focus on culture-specific concepts and curate a dataset to answer the following research questions: *How well do modern VLMs perform in recognizing culture-specific concepts, and can they adapt to these concepts by leveraging multimodal contextual information?*

Prior studies show that Large Language Models (LLMs) (Johnson et al., 2022; Dwivedi et al., 2023), as well as CLIP-style vision encoders (Richards et al., 2024; Nwatu et al., 2023) are biased towards Western cultures (Liu et al., 2021a). Given the paradigm of developing VLMs by combining together pre-trained vision encoders (Radford et al., 2021; Zhai et al., 2023), and LLMs (Dubey et al., 2024; Jiang et al., 2023), recent work (Ananthram et al., 2024) has highlighted that these biases transfer to multimodal models.

Towards more culturally inclusive VLMs, we

[1]Github repository here.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7917–7936
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

develop a *CultuRe-specific Probing Evaluation* (CROPE) dataset. Similar to previous VL probing datasets (Hendricks and Nematzadeh, 2021; Shekhar et al., 2017; Li et al., 2023), we formulate the task as binary questions, which probe for the presence of a concept in the image, as shown in Figure 1. To stress-test a model's knowledge, we construct hard negative questions in which the concept in question and the concept in the image are visually or functionally similar. Although language and culture interact (Hovy and Yang, 2021; Hershcovich et al., 2022), we limit our dataset to English, disentangling cultural from linguistic knowledge.

Following previous work (Neeman et al., 2023), we designed CROPE to evaluate two types of knowledge: (1) *parametric*—knowledge encoded in the model weights, and (2) *contextual*—external knowledge (e.g., a Wikipedia summary and corresponding image) given to the model to describe the culture-specific concept. We experiment with several state-of-the-art open-source and open-weights VLMs with four different conditions where we vary the amount of contextual information (see Figure 1). Our findings illustrate that with no context at all, models exhibit a considerable performance drop relative to common concepts (Li et al., 2023) that are prevalent in most established training data for developing VLMs. Surprisingly, when provided with contextual knowledge, the performance of most models deteriorates even more.

We analyze this behavior by inspecting the performance on an easy version of CROPE, where the target concept and the concept in the question belong to separate categories (e.g., food and beverage vs animals). In this case, most models show a performance improvement when provided with the textual information indicating that models struggle to differentiate between hard negative concepts. Finally, we conduct a human evaluation that highlights which type of context (Wikipedia summary, image, or both) is beneficial for humans when completing the same task. We find that the information provided by the text and the image modality is complementary for humans, which suggests that the observed model performance results from a lack of multimodal context understanding.

2 Related Work

2.1 Cultural Knowledge of VLMs

Evaluation of Cultural Knowledge Previous work has aimed to evaluate the performance of VLMs across cultures and languages. MaRVL (Liu et al., 2021b) tests cross-lingual transfer on visual reasoning with culturally relevant concepts, while GD-VCR (Yin et al., 2021) focuses on commonsense reasoning regarding traditions and events from different regions, as depicted in movie scenes. XM3600 (Thapliyal et al., 2022) and MaXM (Changpinyo et al., 2023), introduce multilingual benchmarks for image-captioning and VQA, respectively, using geographically diverse images from Open Images (Kuznetsova et al., 2020). However, as noted by Shankar et al. (2017), these images do not necessarily feature culture-specific concepts despite their regional diversity.

Concurrent efforts aim to assess the capabilities of VLMs in diverse cultural contexts. These works vary in their focus, from regional traditions to multilingual capabilities. Sea-VQA (Urailertprasert et al., 2024) introduces a dataset for multi-hop reasoning on cultural concepts from eight Southeast Asian countries. GlobalRG (Bhatia et al., 2024) targets geo-diverse image retrieval and visual grounding of culture-specific concepts. CulturalVQA (Nayak et al., 2024) and CVQA (Romero et al., 2024) present knowledge-based questions centered on cultural understanding, with CVQA offering a multilingual component. While these works evaluate broader types of cultural knowledge, our work complements these efforts by isolating the evaluation of the recognition and adaptation to culture-specific concepts. CROPE additionally assesses the ability of VLMs to improve cultural understanding by leveraging non-parametric, multimodal knowledge which could serve as a scalable solution for incorporating underrepresented concepts given the constraints of model size.

Cultural Adaptation Methods Relatively few studies have focused on adapting VLMs to culture-specific concepts. Most prior work has concentrated on encoder-only VLMs proposing approaches such as geo-diverse pretext objectives (Yin et al., 2023), data augmentations through code-switching and image editing (Li and Zhang, 2023a), or interventions in the pretraining data composition (Pouget et al., 2024; Ignat et al., 2024). In this work, we explore whether multimodal contextual knowledge provided to generative VLMs at inference can enhance the understanding of cultural concepts.

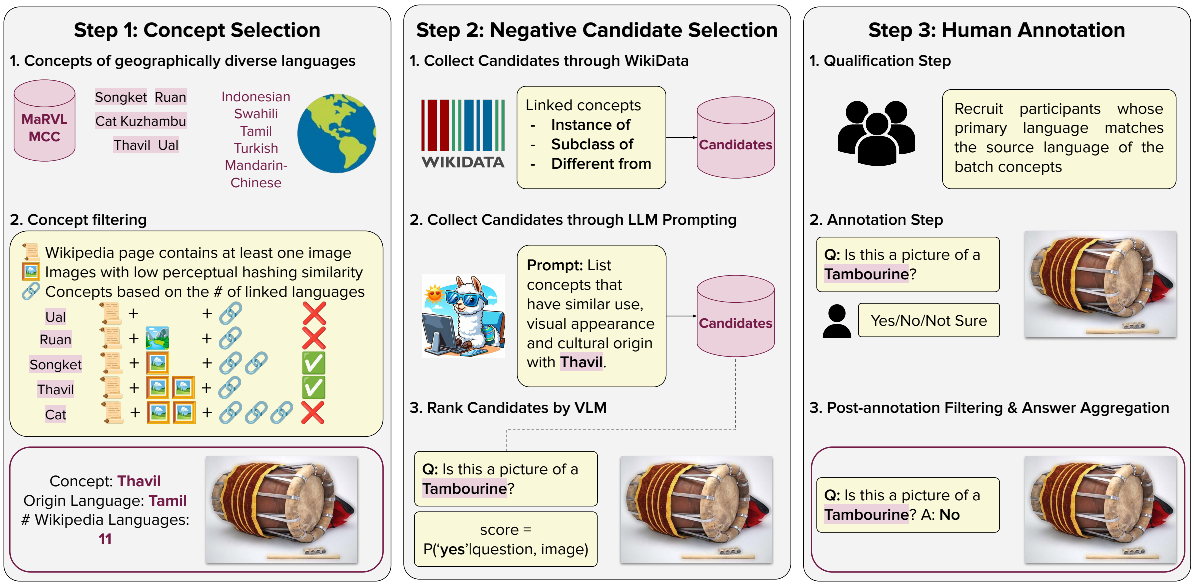

Figure 2: Overview of the dataset creation methodology. We start from a collection of concepts from geographically diverse languages. We collect a pool of challenging negative candidates from Wikidata and by prompting an LLM. Then, we use a VLM to rank candidates and sample up to three candidates per image. To verify each example, we ask human annotators who are proficient in the original concept language and English to annotate the images. Finally, we aggregate the labels and filter out ambiguous examples.

2.2 Multimodal Context in VLMs

Inspired by the in-context learning capabilities of LLMs, modern VLMs (Alayrac et al., 2022; Huang et al., 2023; Laurençon et al., 2024a; McKinzie et al., 2024) have evolved from accepting single-image text pairs to more flexible interleaved image-text inputs. This behavior is driven by training on multimodal web data, which enables VLMs to reason over multimodal documents, compare groups of images, or handle co-references across multimodal contexts (Laurençon et al., 2024; Wang et al., 2024a; Lin et al., 2024; Xue et al., 2024).

To evaluate these capabilities, several recent benchmarks have been proposed. These aim to evaluate the perceptual abilities (Wu et al., 2024; Fu et al., 2024), cross-image reasoning (Li et al., 2024b; Jiang et al., 2024; Li et al., 2024c), or long-context processing (Song et al., 2024) in the presence of distractor images (Sharma et al., 2024; Wang et al., 2024c). Contrary to the existing benchmarks, our work focuses on culture-specific concepts and aims to expand the inclusivity of VLMs through parametric or contextual knowledge.

3 CROPE Dataset

The objective of CROPE is to serve as a challenging evaluation set that probes the capabilities of modern VLMs to recognize and adapt to culture-specific concepts. Figure 2 outlines the steps of the dataset development: 1) concept selection from different cultures, 2) negative candidate selection via model-based sampling, and 3) human annotation that verifies the correctness of each example.

3.1 Dataset Creation Methodology

Concept Selection We use concepts and images collected from the multilingual visual reasoning dataset MaRVL (Liu et al., 2021a), and the follow-up Multimodal Cultural Concepts (MCC) dataset (Li and Zhang, 2023a). Both datasets contain concepts and associated images from five different language origins: Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese. During the collection of these resources, native speakers were asked to provide concepts that are representative of the speaker's culture but common in their everyday experience. As a result, these concepts are not necessarily unique to a particular culture (Cao et al., 2024), but span both universal and culture-specific concepts (Karamolegkou et al., 2024).

We keep the concepts for which we can recover an English Wikipedia page. Using the Wikipedia API, we retrieve the page summary, available images, and their captions. We discard concepts whose Wikipedia page does not contain any images and filter the dataset images based on perceptual hashing similarity[2]. To focus on culture-specific concepts, we use the number of linked Wikipedia pages in different languages as a proxy. Prior work has identified that cultural content is covered in significantly fewer languages compared to general topics (Miquel-Ribé and Laniado, 2018). For example, the page for 'Cat' is available in 267 languages, while for 'Thavil' it appears only in 11 languages. We keep the 40 concepts per language with the least number of linked Wikipedia pages (see Appendix A.1 for details). Lastly, as Wikipedia's concept coverage varies per language, we remove concepts that appear in few languages but are well-represented in image-text datasets.

Negative Candidate Selection We aim to create a pool of negative concepts that stress-test the models' knowledge of a concept. For this purpose, we collect negative concepts from two sources. First, for each concept, we use the Wikidata API[3] to collect concepts that are linked to the target concept either with the 'different from' property or are children of a concept identified by the properties 'subclass' and 'instance of'. For example, 'Thavil' is a subclass of 'Membranophones' from which we retrieve all other subclass musical instruments as possible candidates, such as the 'Tambourine'. Second, we prompt LLama3 (Dubey et al., 2024) to provide a list of 10 concepts that have similar use, visual appearance, and cultural origin with the target object. This process creates a pool of negative candidates for each target concept. Finally, to select challenging concepts, we rank candidates using Paligemma (Beyer et al., 2024). We provide the image of the positive concept, the question 'Is a <negative_concept>?' and measure the probability that the model answers incorrectly. While generating the dataset, we sample up to three negative candidates based on their scores.

Human Annotation We collect ground truth answers through human annotation to 1) minimize the false negatives due to the co-occurrence of multiple concepts in images (e.g., different clothing items) and 2) ensure that the target concept is distinguishable. For each sample, the participants are provided with a definition containing the Wikipedia image and summary for a concept and asked to determine if the concept is present in the second image. In addition to 'Yes' and 'No', the participants can answer 'Not Sure' for cases where the definition is not sufficient to determine the answer.

We recruit participants through the Prolific platform[4]. To ensure familiarity with the concepts depicted in the target image, our pool of participants is limited to those whose primary language matches the concept's origin language and are also fluent in English. We recruit at least 10 participants for each language through a qualification task and collect three annotations per question. We discard samples where at least two participants answered 'Not Sure' or there is no consensus among the annotators. Details are provided in Appendix A.2.

3.2 Dataset Summary

In total, we collected 1060 examples of binary questions, where each example is also accompanied by the Wikipedia summary and images of the concept in question. The annotations of our study show moderate to high inter-annotator agreement (Krippendorff's alpha=0.76). Note that the answer distribution is imbalanced ('Yes': 35.3% , 'No': 64.7%) to probe the knowledge of the target image concepts. We provide supplementary information regarding the dataset in Appendix B.1.

4 Experimental Setup

Models We experiment with a variety of open-source and open-weights generative models up to 11B parameters (see Table 7) that achieve state-of-the-art performance on established VQA benchmarks (Hudson and Manning, 2019; Goyal et al., 2017). In our study, we categorize models based on the following: 1) whether the model has been trained with *multi-image data*—which we expect should benefit from context information the most, 2) whether the model is trained with *multilingual image-text data*—which we expect to show better zero-shot performance on the examined concepts.

Experimental Setup To disentangle the impact of parametric and contextual knowledge (Neeman et al., 2023), our study covers four different experimental conditions: 1) **Zero-shot** (*parametric*), where a model is only given the target image and question; 2) **Textual context**, where the model is also given the Wikipedia summary of the concept as additional context; 3) **Visual context**, where the model is given the Wikipedia image of the con-

[2]Using the python library imagehash.
[3]https://www.wikidata.org/wiki/Wikidata:REST_API
[4]www.prolific.com

Model	MI	ML	POPE F1	CROPE F1	Precision	Recall	Yes %	Consistency
Majority class (No)	-	-	33.33	32.26	50.00	39.22	0	0
LLaVA-1.5 (2024a)	×	×	82.19	62.31	67.15	67.33	60.64	38.38
MOLMO (2024)	×	×	83.88	62.50	70.41	69.35	**66.73**	43.47
LLaVA-NeXT (2024b)	×	×	85.91	64.46	67.33	68.15	55.86	41.43
Phi-3-Vision-128K-Instruct (2024)	×	×	84.56	68.94	70.53	68.98	31.86	40.51
Llama-3.2-Vision-Instruct (2024)	×	×	85.06	**79.11**	**79.38**	**80.43**	42.95	**64.77**
Paligemma (2024)	×	✓	85.57	68.89	70.78	70.97	47.83	46.45
XGen-MM-Interleaved (2024)	✓	×	86.82	69.24	74.30	74.95	61.87	48.86
Idefics2 (2024b)	✓	×	84.13	70.56	75.15	75.50	58.78	52.37
Mantis-Idefics2 (2024)	✓	×	84.13	70.79	72.82	74.31	53.04	52.69
VILA (2024)	✓	×	82.30	74.92	77.55	74.13	28.15	52.07
InternLM-XComposer-2.5 (2024)	✓	+ZH	84.90	64.73	70.46	70.57	63.60	44.01
LLaVA-OneVision (2024a)	✓	+ZH	**87.66**	70.68	75.20	76.17	60.76	53.70
Qwen2-VL-Instruct (2024b)	✓	✓	86.91	74.06	77.56	79.13	58.17	58.53
mPLUG-Owl3 (2024)	✓	+ZH	87.03	74.39	75.19	77.17	49.50	56.44
Gemini (gemini-1.5-flash-latest Sep 2024) (2023)	✓	✓	88.20	78.74	80.11	78.60	38.59	60.12
Gemini (gemini-pro-latest Sep 2024) (2023)	✓	✓	88.45	79.27	87.35	72.81	30.00	50.19
GPT-4o (gpt-4o-2024-08-06) (2023)	✓	✓	88.66	88.87	89.02	88.92	37.58	82.49

Table 1: Zero-shot performance of models on POPE (adversarial split) and CROPE. MI: Model has been trained with interleaved image-text data. ML: Model has been trained with multilingual image-text data. +ZH: Usage of image-text data in Chinese. The performance of closed-source models is indicated in gray.

cept as well as the caption of the image[5]; 4) **Multimodal context**, where the model receives both the summary of the concept and the corresponding exemplar image from Wikipedia. We make a distinction between the 'Multimodal' and 'Visual' context conditions, although the latter technically includes both an image and its caption, as the caption does not provide a definition of the concept. To reduce the effect of prompt sensitivity (Salinas and Morstatter, 2024), we use three prompts for all models and report the average performance. We apply the same evaluation setup for zero-shot results on POPE to ensure a fair comparison. Finally, the conditions providing images as context apply to models that accept interleaved image-text input.

Evaluation Metrics Following POPE (Li et al., 2023), we report the F1-score, precision, recall, as well as the percentage of positive responses. We additionally report the consistency score (Hudson and Manning, 2019), where the model receives +1 for correctly answering 'Yes' and 'No' questions for a given image else 0. To ensure reproducibility, we employ greedy decoding in all experiments.

[5]For images without a caption. we use the template: `An image of <concept>`.

5 Results

5.1 Zero-shot Performance

Table 1 shows the performance of all models in the zero-shot setting. Models score high on the POPE adversarial split that probes for the existence of common and frequently co-occurring objects in images. With the exception of GPT-4o, performance drops substantially on CROPE, which targets more culturally specific objects. Even though Gemini-1.5-Pro achieves the second-highest F1 score, there is still a considerable 9-point gap. As the behavior of proprietary models is difficult to explain and often not reproducible, our remaining analysis focuses on open-source and open-weight models.

We observe that the F1 score of several open models (LLaVA-1.5, LLaVA-NeXT, MOLMO, InternLM-XComposer) drops by up to 20 points when they are evaluated on culture-specific concepts. The high Yes% for these models indicates that they struggle to differentiate the negative candidates from the actual concept in the images. The model with the strongest zero-shot performance is Llama-3.2, outperforming others by a large margin in terms of F1 and Consistency. The advantage of Llama-3.2 could be explained by its extensive

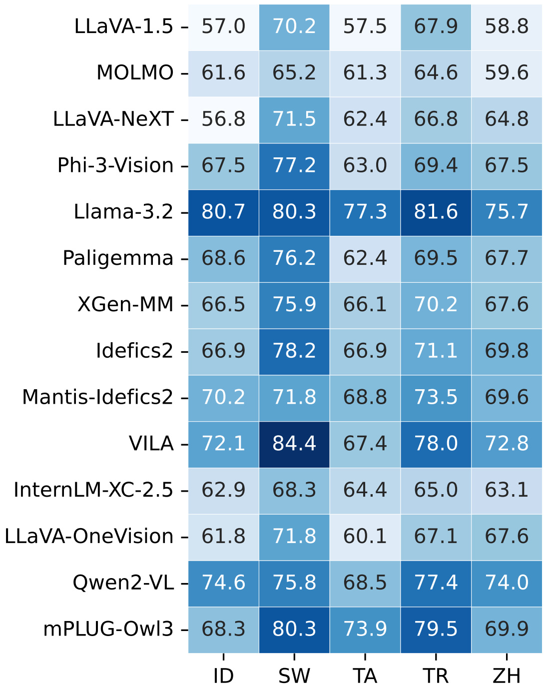

Figure 3: Zero-shot F1 score per source language.

pretraining on a dataset of 6B image-text pairs that underwent thorough preprocessing and deduplication to maximize data diversity (Dubey et al., 2024). Among the four models with the highest Consistency, three remaining are trained with multilingual image-text data. These results are in line with recent work (Pouget et al., 2024) that advocates for including multilingual data in the pretraining mixture, as this can enable maintaining performance on standard English benchmarks while enhancing cultural knowledge.

Figure 3 reports model performance based on the source language of the concepts. Models perform reasonably well on images with concepts sourced from Swahili and Turkish but underperform on concepts from other high (Chinese) or mid-resource[6] (Indonesian, Tamil) languages. It is important to consider that VLMs are built by integrating vision and language backbones through joint training stages with image-text examples. Therefore, the resource characterization of languages supported by multilingual LLMs may not accurately represent the VLM landscape. Building on Pouget et al., we need to systematically analyze the data availability and coverage of cultural concepts across different training stages of generative VLMs.

[6]Following the categorization of (Joshi et al., 2020).

5.2 Performance with Contextual Knowledge

We explore if the performance gap between common and culture-specific concepts can be addressed through contextual knowledge. Figure 4 shows the performance of all models under the different context conditions. While contextual knowledge does not yield notable improvements in any condition, we find that the visual context is more beneficial than the textual for most models, and the multimodal context tends to be the most helpful. We observe that in the Textual context condition, where a concept summary is provided in the prompt, all models show an increase in the percentage of 'Yes' responses, leading to increased false positives. Only three models, XGen-MMm InternLM-XComposer, and Qwen2-VL, exhibit improved performance compared to the zero-shot setting. Nevertheless, the best-performing model-context combination does not surpass the highest zero-shot performance for open-weights VLMs.

Our findings suggest that current VLMs struggle to process multimodal contextual information. We consider two possible reasons for this behavior. First, the models might be sensitive to the task structure, which, due to the concept's summary, includes a relatively lengthy prompt. To test this, we evaluate the same models on an easier version of CROPE, which does not necessitate reasoning about subtle differences (see next paragraph). Second, the contextual information may not suffice to disambiguate the concept in question and in the image. We address this by comparing against human performance in ablated contexts in Section 6.

Performance with easy negative candidates To test the sensitivity of VLMs with regard to the input context, we create an easier version of the dataset by selecting random negative candidates. In particular, we use the class of the target concept to sample negative candidates belonging to a different concept category. This process creates a significantly easier version where the target image depicts a different concept category than the concept in question (e.g., a beverage vs. an animal).

We evaluate models both in zero-shot and Textual conditions. Figure 5 shows the relative performance change between the two conditions for the original and the easy dataset. These results indicate two groupings based on the models' behavior when the length of the input prompt is increased by including the Wikipedia summary. The first group comprises most models that show improved per-

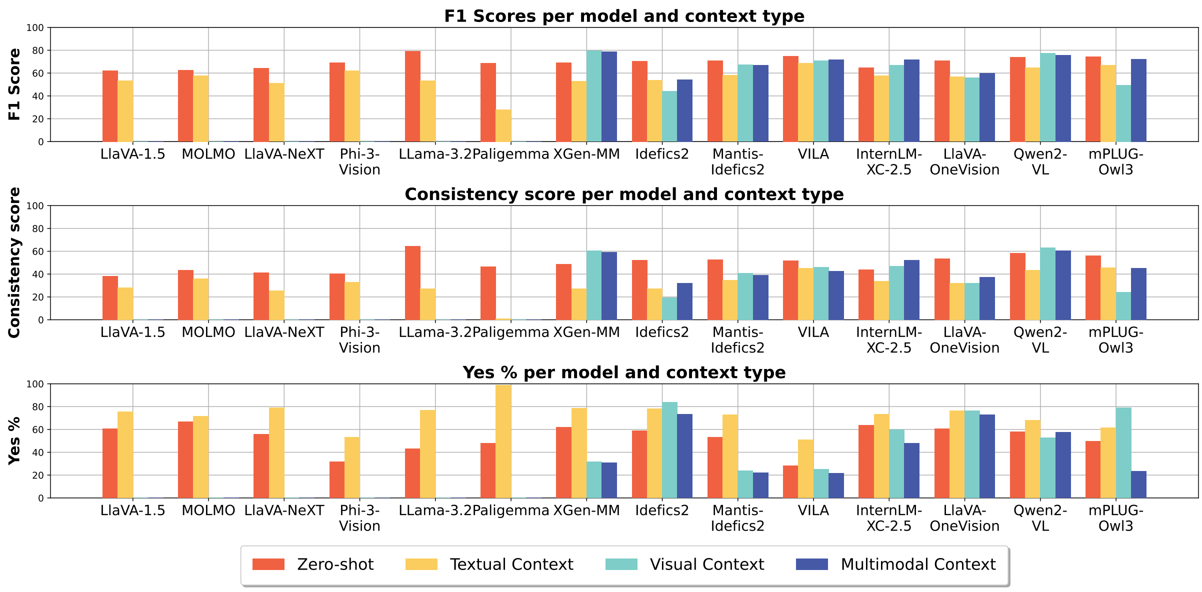

Figure 4: Performance with different context types. All VLMs are negatively impacted when including the concept summary in question (Textual Context). Out of the 7 VLMs that accept multimodal context, only XGEN-MM and InternLM-XComposer benefit from multimodal contextual information.

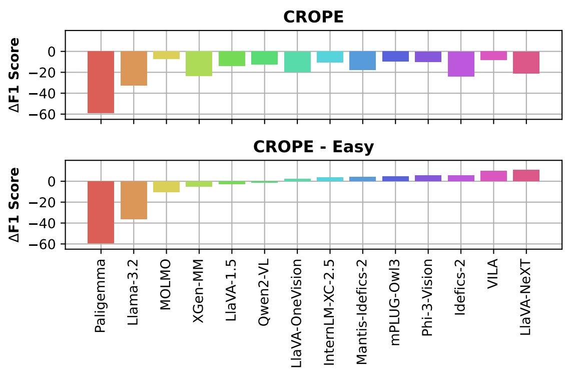

Figure 5: Relative performance of Zero-shot vs Textual conditions for the original (top) and easy (bottom) versions of CROPE. Textual summaries benefit most models when differentiating between easier candidates.

formance or a marginal drop in the easy version. These models seem to be robust to the increased input prompt but struggle to differentiate between similar concepts in CROPE. The second group (Paligemma, Llama-3.2, MOLMO) includes models with comparable relative drops in both datasets. This behavior can be attributed to sensitivity to the longer input resulting from adding the summary.

Performance with increased visual context We also examine the impact of providing more exemplar images in the context of the VLMs. To do so, we keep only the samples with at least three available images and focus on the models that show

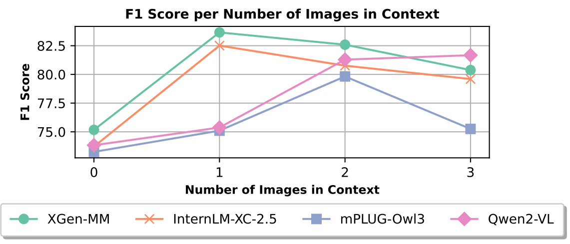

Figure 6: F1 score with varying number of images in the context.

the strongest performance with multimodal context. As shown in Figure 6, increasing the context images from one to two leads to a better F1 score only for mPLUG-OWL3 and Qwen2-VL but has a negative effect on the other models. However, further increasing the context to three images hurts performance for all models except Qwen2-VL. These results align with concurrent studies showing that the performance of highly capable VLMs deteriorates with increased images in the context (Zong et al., 2024; Wang et al., 2024c).

6 Human Evaluation

We conduct a human evaluation across three conditions mirroring those of Section 4: **Textual context**, **Visual context**, and **Multimodal context**. Each described condition relates to the modality by which information is given to participants before their annotation of examples containing unfamiliar con-

	Estimate	SE	z	$p_{\text{val.}}$	$p_{\text{adj.}}$
Intercept	1.51	0.15	9.94	0.000	**0.000**
Textual Context	-0.45	0.20	-2.26	0.024	**0.042**
Visual Context	-0.41	0.20	-2.03	0.042	**0.042**

Table 2: Results of the regression model. $p_{\text{val.}}$: significance of initial findings; $p_{\text{adj.}}$: adjusted $p_{\text{val.}}$ after a Benjamini & Hochberg correction.

Context Type	F1	Precision	Recall	Yes %
Textual	73.12	74.99	73.29	63.14
Visual	74.51	74.66	74.51	54.00
Multimodal	80.92	80.95	80.89	58.29

Table 3: Human performance per context condition.

cepts. Note that we do not target participants from the cultures in our dataset who would be familiar with the image concepts. The aim is not to establish a human baseline for the zero-shot condition, as VLMs are expected to serve users from diverse backgrounds. Instead, we examine how the modality of information influences human judgments to put into perspective the behavior of VLMs under similar contextual settings.

Participants & Stimuli We sample 100 examples with a balanced answer distribution and used power analysis (Cohen, 2013; Lakens and Caldwell, 2021) to ensure that our experiment is sufficiently powered (80%). The design of the study is between-subjects, and we recruit 36 participants per condition, each of whom is given 10 examples. To gauge the overall levels of familiarity of our sample, we asked participants to rate their familiarity with target concepts on a 5-point Likert scale. The median of the ratings across all concepts and conditions is 1 while the 75th percentile is 2, validating that most participants were unfamiliar with the sampled concepts.

Experimental Design We use a mixed-effect regression model, with our dependent variable being the participants' binary response and our predictors matching the three conditions. Since participants annotated multiple examples (Schielzeth et al., 2020; Raudenbush, 1994), the annotator ids were included as random factors. Finally, we evaluate the possible effects of multiple comparisons via a Benjamini-Hochberg correction (Thissen et al., 2002; Benjamini and Hochberg, 1995).

Results The results of the mixed regression model can be seen in Table 2. We report a significant negative effect on both the Textual and the Visual context conditions compared to our baseline (Multimodal). These results indicate that participants found the information presented through a combination of image and text to be significantly more helpful as expressed through more correct responses than when provided through either format alone. This is in contrast with the behavior of VLMs, whose performance decays or, at best, improves minimally with the addition of any form of contextual knowledge. Additionally, Table 3 shows human performance, which is well above random chance, even in the Textual condition. This validates that the summaries and exemplars can help reach a correct answer for unfamiliar concepts.

7 Conclusion

In this work, we introduce CROPE, an evaluation benchmark for probing the parametric and contextual knowledge of VLMs on culture-specific concepts. Our results identify a significant performance disparity of state-of-the-art open VLMs on concepts that appear commonly in VL datasets and CROPE. We also explore whether VLMs can adapt to culture-specific concepts with multimodal contextual information and find that most models fail to utilize this context. We show that this is not necessarily the case when the models are required to compare semantically distant concepts, which indicates that current VLMs struggle to reason about nuanced differences. Conversely, our findings suggest that humans unfamiliar with the concepts in question benefit from multimodal information.

Discussion Our investigation raises the question: *Are modern VLMs truly capable of learning new concepts in-context?* Early work (Tsimpoukelli et al., 2021) showed promise for VLMs capable of fast-mapping, which refers to learning to bind new concepts to images with limited contextual information (Carey and Bartlett, 1978). The authors speculate that the binding capacity of a model can be improved with richer visual or textual support. Our analysis shows that current VLMs have not yet met this expectation, as they exhibit, at best, marginal improvements with relevant context. Thus, processing arbitrary interleaved image-text formats remains a challenge.

Finally, we do not take the position that models should solely rely on non-parametric knowledge to perform well on tasks that require cultural understanding. Given that benchmarks often become quickly saturated (Kiela et al., 2021) and the growing interest in more pluralistic representation in VLMs, we anticipate future model iterations to 'solve' the task in a zero-shot manner. Nevertheless, CROPE provides the opportunity to stress-test the current knowledge of culture-specific concepts embedded into the model weights, as well as the utilization of contextual knowledge. We find that there is plenty of room for improvement across both evaluation axes, and hope that CROPE can contribute towards the development of more capable and culturally inclusive VLMs.

8 Limitations

We build on previous collections of culturally relevant concepts (Liu et al., 2021a) that cover only a small percentage of global cultures. It is important that future work expands this collection to incorporate a broader range of cultural contexts for a more comprehensive evaluation. Moreover, our dataset is limited to English. This does not address any potential disparities in performance across languages (Zhang et al., 2023) or cases where there is no equivalent translation (Majid et al., 2015).

Additionally, our findings about the behavior of VLMs are based on models with up to 11B parameters given relevant information retrieved directly from Wikipedia. We do not examine the impact of scale or address the issue of how to retrieve informative multimodal contexts. This constitutes an active area of research (Wei et al., 2023), as it is an important consideration for practical applications.

9 Ethics Statement

CROPE contains human labels over questions regarding images that depict culture-specific concepts. The data collection has been approved by the ethics and data protection committees of our institution. Prior to completing the task, participants were given an information sheet detailing the purpose of the study, their rights as participants, the compensation provided for participation, and a statement assuring that no personally identifiable information would be included in any report resulting from the study. Each participant then had the option to sign a consent form acknowledging the information provided.

For the data collection, we engaged with participants who were native speakers of the concepts' source languages to ensure inclusivity and a high degree of familiarity with the target concepts during the data collection process. However, recruiting annotators solely based on their native language potentially overlooks minority communities. While we acknowledge the limitations of our current work, we are hopeful that it will contribute to advancing cultural understanding of modern VLMs. We encourage future research to explore additional criteria within cultural groups to facilitate more representative sampling.

Acknowledgements

This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Program at the University of Edinburgh. In addition, the authors acknowledge the use of the HWU high-performance computing facility (DMOG) and associated support services in the completion of this work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. In *Workshop on Common Sense Knowledge Graphs (CSKGs)*.

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

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

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

Lars Borin, Bernard Comrie, and Anju Saxena. 2013. The intercontinental dictionary series–a rich and principled database for language comparison. *Approaches to measuring linguistic differences*, 285:302.

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, and Daniel Hershcovich. 2024. Exploring visual culture awareness in gpt-4v: A comprehensive probing. *arXiv preprint arXiv:2402.06015*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67. Association for Computational Linguistics.

Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development*.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931. Association for Computational Linguistics.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

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

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809. Association for Computational Linguistics.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders

Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

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

Oana Ignat, Longju Bai, Joan C. Nwatu, and Rada Mihalcea. 2024. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1239–1259. ELRA and ICCL.

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*, abs/2310.06825.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *ArXiv*, abs/2203.07785.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Antonia Karamolegkou, Phillip Rust, Ruixiang Cui, Yong Cao, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66. ACL.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Daniël Lakens and Aaron R. Caldwell. 2021. Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920951503.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024a. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models?

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference* *on Computer Vision and Pattern Recognition*, pages 13299–13308.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024c. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics.

Zhi Li and Yin Zhang. 2023a. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276. Association for Computational Linguistics.

Zhi Li and Yin Zhang. 2023b. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024c. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*.

Asifa Majid, Fiona Jordan, and Michael Dunn. 2015. Semantic systems in closely related languages.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Physics*, 6.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070. Association for Computational Linguistics.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702. Association for Computational Linguistics.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *arXiv preprint arXiv:2405.13777*.

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

Stephen W Raudenbush. 1994. Random effects models. *The handbook of research synthesis*, 421(3.6).

Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. 2024. Does progress on object recognition benchmarks improve generalization on crowdsourced, global data? In *The Twelfth International Conference on Learning Representations*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem

Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4629–4651. Association for Computational Linguistics.

Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9):1141–1152.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*.

Aditya Sharma, Michael Saxon, and William Yang Wang. 2024. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. *arXiv preprint arXiv:2406.16851*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. "foil it! find one mismatch between image and language caption". In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 255–265.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788. Association for Computational Linguistics.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729. Association for Computational Linguistics.

David Thissen, Lynne Steinberg, and Daniel Kuang. 2002. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, 27(1):77–83.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. 2024c. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *The Twelfth International Conference on Learning Representations*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10961.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. In *EMNLP*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *arXiv preprint arXiv:2403.13164*.

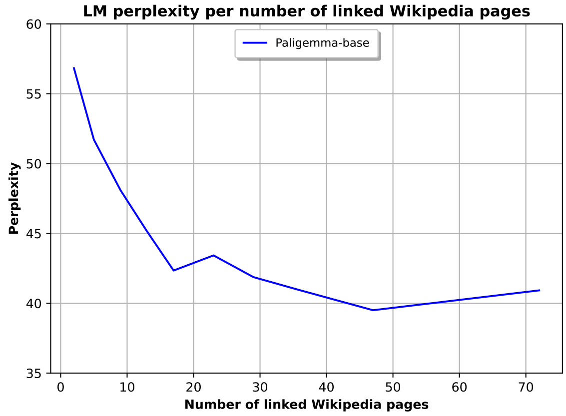

Figure 7: Perplexity of Wikipedia image captions vs the number of linked languages for the page including that image.

A Dataset Collection

A.1 Model Perplexity vs Number of Linked Languages

During our dataset construction, we focused on more culture-specific concepts by filtering out concepts based on the number of languages in Wikipedia. Intuitively, if a concept is not present on Wikipedia for a particular language, it is likely that the concept is less common. This is particularly impactful in the case of VLMs (Laurençon et al., 2024; Deitke et al., 2024) and their respective backbone LLMs (Groeneveld et al., 2024; Soldaini et al., 2024) where Wikipedia is often considered a high data resource.

To showcase this, we measured the perplexity of Paligemma when completing captions from images of Wikipedia containing concepts with varying number of links. Figure 7 shows the perplexity aggregated for different number of linked languages in Wikipedia. We observe a clear trend, where the more a concept is represented in multiple languages, the lower the perplexity of models.

A.2 Human Annotation

Both the Data Annotation and Human Evaluation were conducted through the Prolific platform. We developed the annotation interface shown in Figure 8 using Gradio[7]. For each sample, participants are given an example image and the Wikipedia summary for a concept and are asked to answer if the concept appears in a second image. For the concepts which are associated with multiple images,

[7]https://www.gradio.app/

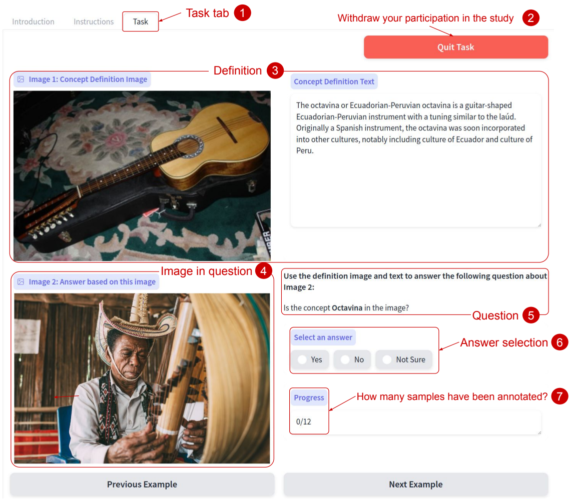

Figure 8: Annotated layout of the annotation interface as shown to the annotators.

Your task is to annotate a series of examples by determining whether a concept is present in an image.
Each concept will be explained with a definition that includes an image and an extract from Wikipedia.

Steps to Follow

1. Click the **Start the task** button at the end of this page to start the annotations. This will be enabled only if you have given consent in the *Introduction* page.
2. Carefully consider the definition image and text provided for each concept. *Try not to rely on your outside knowledge* and pay attention to the details, as there can be small differences that distinguish two concepts.
3. Answer whether the concept is present in the second image. You can answer with one of the following:
   - "Yes" if the concept is present.
   - "No" if the concept is not present.
   - "Not Sure" if you are uncertain. Note that frequent use of "Not Sure" may indicate a lack of attention and will be manually reviewed.
4. Click "Next" to move to the next example.
5. Ensure both the definition image and the example image have changed before making your annotation.
6. When you have completed all annotations, a "Complete the task" button will become available. Clicking it, will take you to the "Completion" tab where you will be able to find the link to return to Prolific and submit your work.

Figure 9: Human annotation instructions.

we manually select the example image to ensure it is representative of the target concept. Participants are compensated with 15.00$/ per hour.

We split our data into batches where each batch consists of examples from the same source language. We recruited participants through a Qualification Stage with the criteria that their primary language is the source language of the batch and that they are fluent in English. Participants signed a virtual consent with details about the data collection, how the data would be used and how they could withdraw if they wished to do so. During the Qualification stage, we asked participants to annotate 10 samples for which the ground truth was known. At the end of the task, we provided participants with corrections for any mistakes and explanations for the correct answer. To ensure high-quality annotations, we only invite participants who answered at least 70% of the qualification examples correctly.

Consequently, for the Annotation Stage, we recruit 10 participants per language (with the exception of Tamil for which we recruit 12). Participants are then asked to annotate the dataset examples. Each example is annotated by three participants. In this task, we also allowed participants to answer with 'Not Sure,' which allows us to discard any ambiguous examples (indicated by at least two 'Not Sure' or tie across the three possible labels). Each participant annotated, on average, 64 samples.

The Human Evaluation study (Section 6) is set up in a similar way. For each condition, we recruit 35 new participants who are fluent in English as the only requirement. Additionally, we ask each participant to annotate 10 samples from different concepts. Before being shown each sample, participants are also asked to rank on a 5-point Likert scale how familiar they are with the concept that appears in the question. Finally, we only allow participants to answer with 'Yes' and 'No'.

B Dataset

B.1 Dataset Analysis

Source Language Code	ID	SW	TA	TR	ZH
# Samples	228	194	230	207	201
# Image Concepts	34	39	24	27	34

Table 4: Number of samples and image concepts per source language.

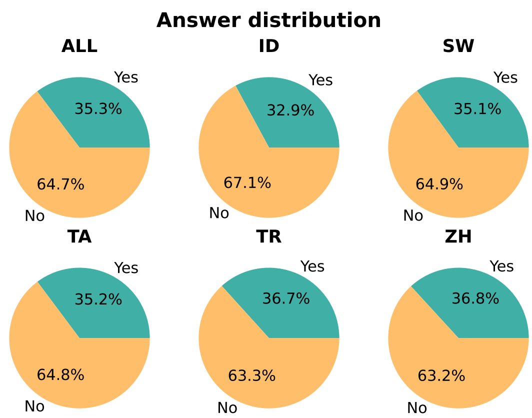

Figure 10: Answer distribution.

The CROPE dataset consists of 1060 evaluation samples with concepts originating from speakers of five typologically diverse languages (Liu et al., 2021b), specifically 228 from Indonesian, 194 from Swahili, 230 from Tamil, 207 from Turkish, and

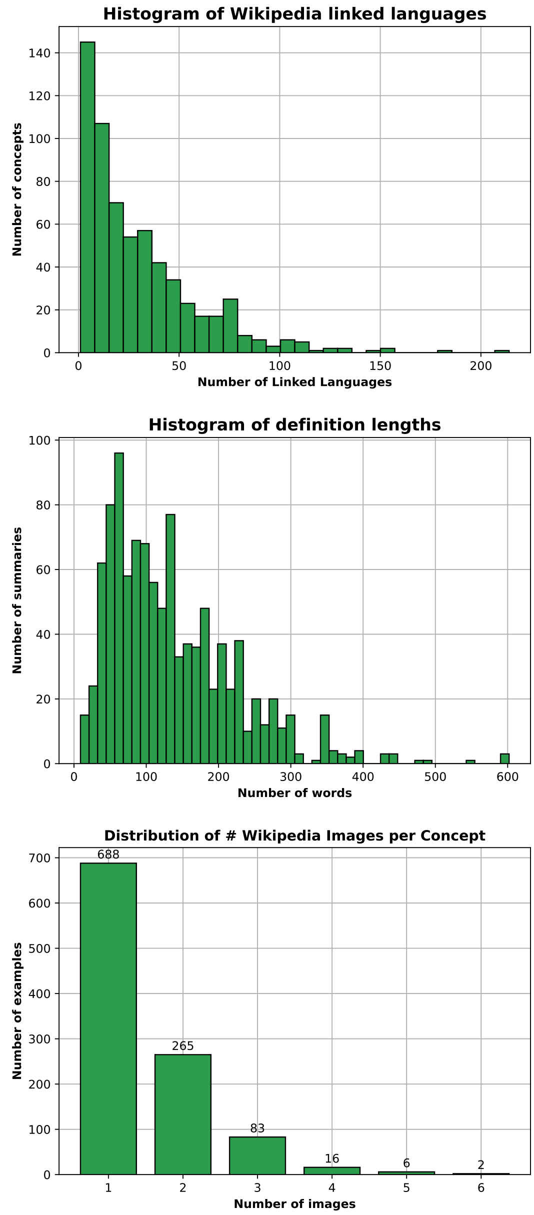

Figure 11: Statistics for the context of the question concepts. (Top) Histogram of the linked languages in Wikipedia for the concepts in CROPE. The majority of the concepts appear in a limited number of languages indicating cultural specificity. (Middle) Histogram of the summary lengths measured as the number of words separated by whitespace. (Bottom) Distribution of # of Wikipedia images per concept.

201 from Chinese. Figure 10 illustrates the answer distribution per source language for each concept. The dataset is imbalanced, with the majority of the ground truth answers being ‘No’. This is done to probe the understanding of the image concept. The answer distribution is maintained across all source languages in our dataset.

The top part of Figure 11 illustrates the number of Wikipedia pages, each with a unique language, that is available per question concept (which includes both image and negative concepts). Note that the vast majority of the concepts that we have selected appear on Wikipedia in less than 50 languages. Additionally, the middle part of Figure 11 shows the distribution of the number of words (separated by whitespace) of the concept summary from Wikipedia. All summaries can easily fit within the context window of modern VLMs (Li et al., 2024a; Laurençon et al., 2024). Finally, the bottom part of Figure 11 depicts the distribution of the number of context images available per concept.

What information does the Wikipedia summary contain that can be used by models to infer the visual appearance of a concept? Many Wikipedia summaries for a concept may contain information that is not particularly relevant to its visual appearance. To quantify the information within the summaries that is useful for identifying a concept, we use GPT-4o (Achiam et al., 2023) to extract the text spans from the summary that can be used to infer a plausible visual appearance of a concept. More specifically, we give the model demonstrations of extracted text-spans from Wikipedia summaries that provide cues for the concept’s visual appearance. We then ask the model to extract the spans for the remaining summaries for all concepts in our dataset. Finally, we compute a word-level overlap between the visually pertinent spans and the entire summary of the concept.

Figure 12 illustrates the histogram of the word-level overlap between the text spans extracted via GPT-4o and the full Wikipedia summary for a concept. We observe that a significant portion of the summary can be used to infer visual properties of a concept, with even some summaries being identified as fully visually relevant by GPT-4o. From manual inspection of the extracted spans, these correspond to relatively short summaries that provide the name, the origin, and the functionality of the concept (e.g, `Chanakhi is a traditional Georgian dish of lamb stew with tomatoes, aubergines, potatoes, greens, and garlic.`).

B.2 Dataset Examples

Table 10 illustrates examples of input-output pairs, including the text and image context. Some of the examples (Example 2 and 4), depict cases where image or the text context is sufficient to answer the

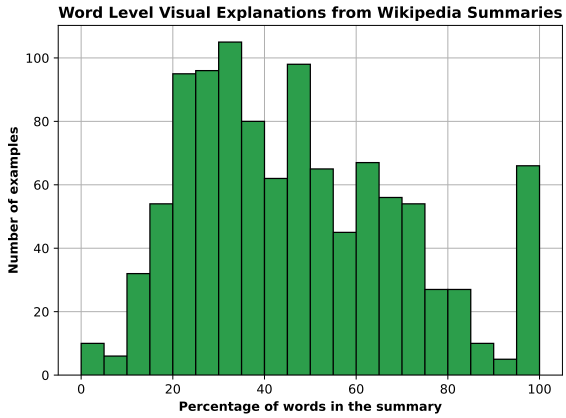

Figure 12: Histogram of word-level overlap percentage between visually relevant text spans and Wikipedia summaries.

question. Other examples in the table (Example 3 and 5), show cases where the information from a single modality do not provide sufficient cues.

B.3 License

We will release the questions and annotations under the CC BY-NC 4.0 license. The licenses for the different resources used for our dataset are as follows: Wikipedia texts and images are co-licensed under CC BY-SA 4.0 and the GNU Free Documentation License. The MaRVL (Liu et al., 2021a) text data are released under the CC BY 4.0 license, and all included images are also CC-licensed. The MCC dataset (Li and Zhang, 2023b) is available on Hugging Face datasets[8] under Apache license 2.0. All data are provided for non-commercial, research-only purposes.

C Implementation Details

Table 7 reports the model tags from Hugging Face[9] for the open-source/open-weights models. Table 8 shows the prompts used in this study. Note that we also follow the guidelines for each model to include system or chat prompts. Experiments were run on an NVIDIA A40 (40GB), requiring approximately 80 GPU hours to obtain the full results.

When the context includes an example image, and there are multiple available images for the target concept, we use the same image that was selected for the human data collection (Appendix A.2). For experiments with a varying number of k of images in context, we always select the first k images and run the experiments for up to three permutations.

[8] https://huggingface.co/datasets

[9] https://huggingface.co/models

Model	F1 Score $\notin$ MMDU	$\in$ MMDU
mPLUG-Owl3	68.62	68.60
InternLM-XC-2.5	73.52	70.02
XGEN-MM	80.33	84.83

Table 5: F1 score under Multimodal context grouped depending on whether the concept is found in the MMDU training data.

D Further Results

Multimodal Performance Analysis We notice that three of the models with the strongest relative performance when using multimodal context (XGen-MM, InternLM-XComposer-2.5, mPLUG-Owl3) include the MMDU dataset (Liu et al., 2024c) in their supervised finetuning data mixture. MMDU is a recent multi-image instruction following dataset with images from Wikipedia, and dialogs generated by prompting GPT-4o with the relevant images and text information. We find that only 19.5% of the concepts in the questions of CROPE appear in the training set of MMDU. Table 5 shows the F1 score under the multimodal condition aggregated based on whether the image concept appears in MMDU. We find no evidence of consistent benefit in MMDU concepts, which indicates that the results in the multimodal condition are not due to pure memorization.

Performance of Larger Models Table 6 shows the zero-shot F1 score on the POPE and CROPE test sets as model parameter size increases. We additionally provide results for InternVL-2 (Chen et al., 2024), which is available across multiple sizes. We observe that while for POPE, performance remains consistent with scale, in CROPE, the F1 score improves. However, only Llama-3.2 (90B) manages to close the gap between culture-specific and common concepts. This is consistent with findings from prior work (Kandpal et al., 2023), which show that while scaling up improves knowledge of long-tail knowledge, substantial gaps remain, particularly for cases with limited support in pretraining data.

Prompt Sensitivity Table 9 shows the mean and standard deviation of the F1 scores for different prompts under each condition.

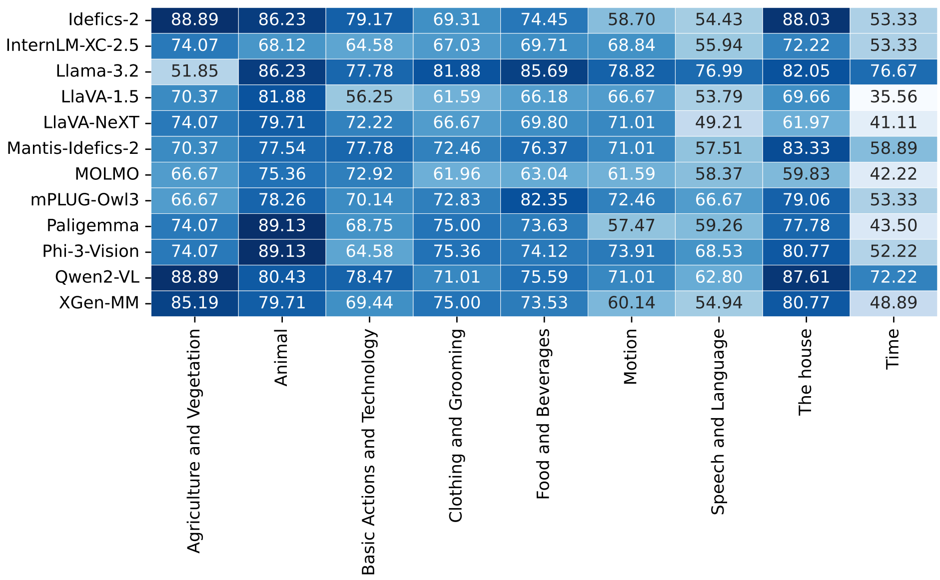

Figure 13: Zero-shot F1 per concept chapter.

Model	POPE	CROPE
Llama-3.2-9B	85.06	79.11
Llama-3.2-90B	85.60	86.32
LLaVA-OneVision-7B	87.66	70.68
LLaVA-OneVision-72B	85.69	77.83
MOLMO-7B	83.88	62.50
MOLMO-72B	84.70	71.41
Qwen2-VL-Instruct-7B	86.91	74.06
Qwen2-VL-Instruct-72B	86.17	77.83

Table 6: Zero-shot F1 Score as model size increases.

Performance per Chapter Each concept is associated with a chapter from the Intercontinental Dictionary Series (Borin et al., 2013). Figure 13 shows the F1 score per concept chapter. We observe that chapters ‘Time’, that includes concepts about celebrations, and ‘Speech and Language’, which includes concepts about musical instruments and visual art forms, are the most challenging across models. On the other hand, most models score highly on ‘Agriculture and Vegetation’, ‘Animal’, and ‘The house’. Overall, we find that different models have different areas of strength and weakness.

E Acknowledgements

We acknowledge the use of GitHub Copilot[10] in the implementation of our research. All final code is verified by the authors. We also acknowledge the use of ChatGPT[11] in improving the clarity of the writing of this paper.

[10] https://github.com/features/copilot

[11] https://chatgpt.com/

Model	Hugging Face Model Name
Idefics2 (Laurençon et al., 2024b)	`HuggingFaceM4/idefics2-8b`
InternLM-XComposer-2.5 (Zhang et al., 2024)	`internlm/internlm-xcomposer2d5-7b`
Llama-3.2 (Meta, 2024)	`meta-llama/Llama-3.2-9B-Vision-Instruct`
LLaVA-1.5 (Liu et al., 2024a)	`llava-hf/llava-1.5-7b-hf`
LLaVA-1.5-13B (Liu et al., 2024a)	`llava-hf/llava-1.5-13b-hf`
LLaVA-Next (Liu et al., 2024b)	`llava-hf/llava-v1.6-mistral-7b-hf`
LLaVA-Next-Vicuna-7B (Liu et al., 2024b)	`llava-hf/llava-v1.6-vicuna-7b-hf`
LLaVA-Next-Vicuna-13B (Liu et al., 2024b)	`llava-hf/llava-v1.6-vicuna-13b-hf`
LLaVA-OneVision (Li et al., 2024a)	`lmms-lab/llava-onevision-qwen2-7b-ov`
Mantis-Idefics2 (Jiang et al., 2024)	`TIGER-Lab/Mantis-8B-Idefics2`
MOLMO (Deitke et al., 2024)	`allenai/Molmo-7B-O-0924`
mPLUG-Owl3 (Ye et al., 2024)	`mPLUG/mPLUG-Owl3-7B-240728`
Paligemma (Beyer et al., 2024)	`google/paligemma-3b-mix-224`
Phi-3-Vision (Abdin et al., 2024)	`microsoft/Phi-3-vision-instruct`
Qwen2-VL (Wang et al., 2024b)	`Qwen/Qwen2-VL-7B-Instruct`
VILA (Lin et al., 2024)	`Efficient-Large-Model/Llama-3-VILA1.5-8B`
XGen-MM (Xue et al., 2024)	`Salesforce/xgen-mm-phi3-mini-instruct-interleave-r-v1.5`

Table 7: Model details: Hugging Face model names.

Prompt Templates
`Answer with yes or no: <QUESTION>`
`<QUESTION>\nAnswer the question using a single word or phrase.`
`Look carefully at the previous image and answer the following question with yes or no: <QUESTION>`

Table 8: Prompt templates. We follow the release guidelines for each model and include the system prompt or chat template as specified. In the Textual context condition, we combine the summary with the question as: `<SUMMARY>\n<QUESTION>`.

Model	Zero-Shot	Textual Context	Visual Context	Multimodal Context
Idefics2-8B (Mistral-7B)	70.56 ± 3.98	53.72 ± 1.16	44.11 ± 6.63	54.31 ± 4.82
InternLM-XComposer-2.5 (InternLM2-7B)	64.73 ± 2.87	57.76 ± 1.92	66.85 ± 1.47	71.94 ± 0.72
Llama-3.2-Vision-Instruct-11B (Llama-3.1-8B)	79.11 ± 2.77	53.22 ± 8.52	-	-
LLaVA-1.5-7B (Vicuna-7B-v1.5)	62.31 ± 2.56	53.55 ± 2.96	-	-
LLaVA-1.5-13B (Vicuna-13B-v1.5)	61.57 ± 2.23	52.88 ± 2.55	-	-
LLaVA-NeXT-Mistral-7B (Mistral-7B-Instruct-v0.2)	64.46 ± 0.79	50.95 ± 2.82	-	-
LLaVA-NeXT-Vicuna-7B (Vicuna-1.5-7B)	64.16 ± 1.47	58.37 ± 4.61	-	-
LLaVA-NeXT-Vicuna-13B (Vicuna-1.5-13B)	63.88 ± 2.46	46.77 ± 0.39	-	-
LLaVA-OneVision-7B (Qwen2-7B)	70.68 ± 2.36	56.82 ± 2.11	55.99 ± 4.26	59.84 ± 4.84
Mantis 8B-Idefics2 (Mistral-7B)	70.79 ± 2.31	58.20 ± 1.98	67.38 ± 1.21	66.76 ± 2.42
MOLMO-7B (OLMo-7B-1124)	62.50 ± 4.97	57.93 ± 4.59	-	-
mPLUG-Owl3 (Qwen2-7B)	74.39 ± 1.38	67.10 ± 2.44	49.54 ± 7.94	72.36 ± 1.00
Paligemma-3B-mix-224 (Gemma-2B)	68.89 ± 3.35	28.10 ± 0.45	-	-
Phi-3-Vision-128K-Instruct (Phi3-mini)	68.94 ± 2.25	62.06 ± 0.92	-	-
Qwen2-VL-Instruct-7B (Qwen2-7B)	74.06 ± 1.58	64.73 ± 2.64	77.25 ± 2.77	75.82 ± 3.58
VILA (Llama-3-8B)	74.92 ± 2.32	68.73 ± 1.03	70.77 ± 3.87	71.60 ± 3.15
XGen-MM-Interleaved (Phi3-mini)	69.24 ± 0.16	52.82 ± 0.56	79.52 ± 0.68	78.75 ± 0.65

Table 9: Mean and standard deviation of F1 score for different prompts.

Exemplar Images	Wikipedia Summary	Target Image	QA and Metadata
	Noken is a traditional Papuan multifunctional knotted or woven bag native to the Western New Guinea region, Indonesia. Its distinctive usage, which involves being hung from the head, is traditionally used to carry various goods, and also children.		Q: Is there a noken in the image? A: Yes Noken ID Clothing and Grooming
 	Injera is a sour fermented pancake-like flatbread with a slightly spongy texture, traditionally made of teff flour. In Ethiopia and Eritrea, injera is a staple. Injera is central to the dining process in Amhara community, like bread or rice elsewhere and is usually stored in the mesob.		Q: Is there injera in the image? A: No Ugali SW Food and Beverages
	Oom-pah , Oompah or Umpapa is an onomatopoeic term describing the rhythmic sound of a deep brass instrument in combination with the response of other instruments or registers in a band, a form of background ostinato...		Q: Is the image about oom-pah? A: No Parai TA Speech and Language
	Shibori is a Japanese manual tie-dyeing technique, which produces a number of different patterns on fabric.		Q: Is this an example of paper marbling? A: No Paper marbling TR Speech and Language
	A siheyuan is a historical type of residence that was commonly found throughout China, most famously in Beijing and rural Shanxi... remaining siheyuan are often still used as subdivided housing complexes, although many lack modern amenities.		Q: Is there a siheyuan in the image? A: Yes Siheyuan ZH The house

Table 10: Dataset examples: Question Concept , Target Concept , Source Language of Target Concept , Chapter . Example 1 (Noken) shows an instance where the textual context complements the image by specifying how the bag is usually worn. Examples 2 and 4 show instances where either the image or the text would be sufficient to answer the question. Example 3 shows an instance where the image exemplar is not as informative, but the text clarifies the type of instrument. Example 5 shows an instance where the text alone does not provide sufficient visual cues.