CVLUE: A New Benchmark Dataset for Chinese Vision-Language Understanding Evaluation

Yuxuan Wang[1], Yijun Liu[2], Fei Yu[1], Chen Huang[1], Kexin Li[1], Zhiguo Wan[1], Wanxiang Che[2], Hongyang Chen[1*]

[1]Zhejiang Lab, Hangzhou, 311121
[2]Harbin Institute of Technology, Harbin, 150001
{yxwang, yufei, huangc, likx, wanzhiguo, hongyang}@zhejianglab.org
{yijunliu, car}@ir.hit.edu.cn

Abstract

Despite the rapid development of Chinese vision-language models (VLMs), most existing Chinese vision-language (VL) datasets are constructed on Western-centric images from existing English VL datasets. The cultural bias in the images makes these datasets unsuitable for evaluating VLMs in Chinese culture. To remedy this issue, we present a new Chinese Vision-Language Understanding Evaluation (CVLUE) benchmark dataset, where the selection of object categories and images is entirely driven by Chinese native speakers, ensuring that the source images are representative of Chinese culture. The benchmark contains four distinct VL tasks ranging from image-text retrieval to visual question answering, visual grounding and visual dialogue. We present a detailed statistical analysis of CVLUE and provide a baseline performance analysis with several open-source multilingual VLMs on CVLUE and its English counterparts to reveal their performance gap between English and Chinese. Our in-depth category-level analysis reveals a lack of Chinese cultural knowledge in existing VLMs. We also find that fine-tuning on Chinese culture-related VL datasets effectively enhances VLMs' understanding of Chinese culture.

Datasets — https://github.com/WangYuxuan93/CVLUE

Introduction

Over the last few years, vision-language pre-training (VLP), as a thriving field, has been drawing extensive attention (Lu et al. 2019; Chen et al. 2020; Cho et al. 2021; Li et al. 2021), leading to significant performance boosts across many VL tasks. It cannot be neglected that the abundance of VL datasets covering various distinct VL tasks (Young et al. 2014; Kazemzadeh et al. 2014; Antol et al. 2015; Chen et al. 2015; Mao et al. 2016; Das et al. 2017; Goyal et al. 2017) plays an essential role in the rapid evolvement of VLMs. However, most of the existing VL datasets are in English. A majority of these datasets, such as NLVR2 (Suhr et al. 2019) and MS-COCO (Lin et al. 2014), are built on top of a hierarchy of concepts selected from English WordNet (Miller 1992), resulting in source images with a North American or Western European bias (Liu et al. 2021). Beyond the English language and Western cultures where these datasets were created, evidence suggests that both the origin (DeVries et al. 2019) and content (Stock and Cissé 2018) of such data are skewed.

Ben.	Lan.	ITR	VQA	VG	VD	VR	IG
VLUE	En.	✓	✓	✓		✓	
CLiMB	En.		✓			✓	
MUGE	Ch.	✓					✓
Zero	Ch.	✓					
CVLUE	Ch.	✓	✓	✓	✓		

Table 1: Tasks included in CVLUE, VLUE, CLiMB, MUGE and Zero. Ben. and Lan. denote Benchmark and Language, respectively. En. and Ch. stand for English and Chinese respectively.

Recently, the community has begun to recognize the importance of cultural differences in large language models (LLMs). Some work has explored the varied performance of LLMs across different cultural contexts (Wang et al. 2023; Li et al. 2024), while other efforts have focused on creating culturally relevant LLM benchmarks (Zhao et al. 2024; Rao et al. 2024). Additionally, there are a few studies investigating cultural awareness in VLMs (Burda-Lassen et al. 2024) and developing multicultural visual question answering (Romero et al. 2024) and visual language reasoning (Liu et al. 2021) datasets. However, these datasets often prioritize coverage of different cultures, with limited task categories and data volumes specific to Chinese culture.

In this work, we focus on the *evaluation of VLMs in Chinese culture, meaning that not only are the texts in Chinese but, more importantly, the images are representative of Chinese culture*. Over the last two years, a significant number of multimodal datasets for Chinese VLM pre-training have been presented (Zhan et al. 2021; Lin et al. 2021; Gu et al. 2022; Liu et al. 2022). However, the development of the benchmark for Chinese VLM evaluation is lagging behind. Many existing Chinese VL datasets exploit images from English VL datasets containing the abovementioned bias.

Some of them, such as Flickr30K-CN (Lan, Li, and Dong 2017), are constructed by translating texts in English VL datasets into Chinese. Others, such as FM-IQA (Gao et al. 2015), Flickr8K-CN (Li et al. 2016) and COCO-CN (Li et al. 2019), are constructed by re-annotating images from English VL datasets in Chinese. Recently, several new datasets with

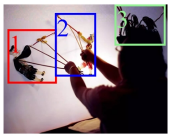
*Corresponding author

1. Two hotpots and some hotpot ingredients are on the table
2. Food and utensils around the hotpots cover the table
3. The two hotpots are boiling, with meatballs and tofu inside
4. Among the food, there are two white hotpots cooking vegetables and meat
5. Around the boiling hotpots, there are meat, vegetables, bowls, chopsticks, and glasses

(a) Image Text Retrieval

Q: What crops are on both sides of the crowd?
A: Rice
Q: Are there more people in the dancing dragon or in the dancing lion?
A: Dancing dragon
Q: How many fewer dancing dragons are there compared to dancing lions?
A: 1

(b) Visual Question Answering



1. The shadow puppet manipulated by the person's left hand
2. The shadow puppet manipulated by the person's right hand
3. The shadow puppet not manipulated by anyone

(c) Visual Grounding

Caption: A man wearing sunglasses is sitting at the front of a dragon boat, directing
Q1: How many people are on the dragon boat?
A1: There are about twenty people on the dragon boat
……
Q10: What color is the dragon boat?
A10: The body of the boat is blue, the bow and stern are yellow and red

(d) Visual Dialogue

Figure 1: Examples of the images and their annotation translated to English for the four tasks in CVLUE.

images collected using Chinese queries on search engines have been presented. However, they are limited to single types of tasks like visual question answering (Wang et al. 2022) or image-text retrieval (Xie et al. 2022).

Chinese is linguistically distinct from English and many other languages, and its speakers comprise one-fourth of the world's population. This necessitates a benchmark dataset specifically designed for Chinese vision-language understanding (VLU). To remedy this issue, we present CVLUE, a new Chinese VL benchmark dataset. We start by selecting categories representative of Chinese culture and manually collect all the images from the Chinese Internet, ensuring that *the source images are commonly seen or representative in the Chinese-speaking population*. The comparison between CVLUE and existing VL benchmark datasets is shown in Table 1.[1] The visual reasoning (VR) task is included in the two English benchmark datasets VLUE (Zhou et al. 2022) and CLiMB (Srinivasan et al. 2022) but not included in any of the Chinese ones. The image generation (IG) task is only included by MUGE, which mainly contains simple iconic images collected from e-commerce platforms and encyclopedias. On the contrary, images in our benchmark were mostly non-iconic ones. The other Chinese dataset Zero (Xie et al. 2022) only focuses on image-text matching and retrieval and comprises five subtasks of a similar type. Our benchmark, by contrast, contains four distinct VL tasks: image-text retrieval (ITR), visual question answering (VQA), visual grounding (VG) and visual dialogue (VD), which evaluate VLMs in Chinese culture from multiple aspects. Examples of images and annotation for the four tasks are shown in Figure 1. See the Appendix for more.

We benchmark several popular open-source multilingual VLMs on CVLUE and established English VL datasets to assess their visual language understanding (VLU) capabilities in both Chinese and English. Furthermore, our in-depth analysis reveals the lack of Chinese culture-related knowledge in existing VLMs. We believe this dataset offers a fair and convenient platform for evaluating VLMs in the context of Chinese culture.

[1]We only compare with benchmarks containing at least two subtasks here.

Related Work

Over the last decade, English VL datasets have experienced rapid development, starting from the most fundamental task of image captioning. Following the popular MS-COCO (Lin et al. 2014) and Flickr30K (Young et al. 2014) datasets, a significant number of VL datasets covering various tasks of visual question answering (Antol et al. 2015; Goyal et al. 2017), visual grounding (Kazemzadeh et al. 2014; Mao et al. 2016), visual entailment (Xie et al. 2019), visual dialogue (Das et al. 2017), etc. have emerged. Recently, an increasing number of English VL benchmarks aiming at different goals have been proposed (Parcalabescu et al. 2022; Zhou et al. 2022; Zheng et al. 2022; Srinivasan et al. 2022), which significantly facilitates the evaluation and comparison of VLMs in English.

Beyond the VL datasets in English, MS-COCO was extended with captions translated to or newly written in German and French (Rajendran et al. 2016), Japanese (Yoshikawa, Shigeto, and Takeuchi 2017) and Chinese (Li et al. 2019). All these datasets exploit images crowdsourced from North America and Western Europe. Researches suggest that they suffer from cultural bias, which may lead to essential limitations for the application in many languages and cultures (Stock and Cissé 2018; DeVries et al. 2019; Liu et al. 2021). In recent years, the community has begun to notice the performance differences of existing VLMs in different cultural applications (Burda-Lassen et al.

2024) and has started to develop multicultural visual question answering (Romero et al. 2024) and visual language reasoning (Liu et al. 2021) datasets. However, these datasets focus on broad cultural coverage, resulting in limited task types and data volume for Chinese.

Over the last two years, an increasing number of Chinese multimodal datasets in the form of image-text pairs have been presented (Lin et al. 2021; Gu et al. 2022; Liu et al. 2022), which has dramatically promoted the evolvement of Chinese VLMs. However, the development of the benchmark dataset for VLM evaluation in Chinese is lagging behind. A great number of existing Chinese VL datasets were constructed by extending English VL datasets with translated (Lan, Li, and Dong 2017) or newly written (Gao et al. 2015; Li et al. 2016, 2019) annotation in Chinese. Wu et al. (2017) presented a Chinese image captioning dataset AIC-ICC, whose images were newly collected from search engines. Recently, two Chinese VQA datasets were introduced, both constructed with newly collected images (Qi et al. 2022; Wang et al. 2022). However, these datasets are limited to single types of tasks and thus insufficient for the comprehensive evaluation of VLMs.

Due to the abundance of English VL datasets, recent English VL benchmarks were mainly constructed using existing datasets. However, the issues discussed above with existing Chinese VL datasets make building a benchmark specifically for Chinese much more challenging. Recently, Xie et al. (2022) introduced a new Chinese VL dataset Zero covering five subtasks. However, all of them involve image-text retrieval/matching and are, therefore, not comprehensive enough to evaluate the general capability of VLMs. Liu et al. (2023) proposed a bilingual VL benchmark MM-Bench, which is first annotated in English and then translated to Chinese using GPT-4. Interestingly, they also released CCBench, a 510-example multiple-choice question answering test set with images closely related to Chinese culture. While it aligns most closely with the goals of this paper, it has significantly less diversity in task types and annotated data than CVLUE.

CVLUE

CVLUE consists of four distinct VL tasks that evaluate a model's capability in Chinese VLU from multiple aspects. The data splits and evaluation metrics are summarized in Table 2. In this section, we describe the procedure we devised for image collection and dataset annotation.

Task	|Train|	|Valid|	|Test|	Metrics
ITR	17,920	3,116	8,973	R@k
VQA	14,362	2,571	7,169	Acc
VG	10,769	1,965	5,385	IoU
VD	3,975	651	2,036	R@k

Table 2: Data splits (in terms of image numbers) and evaluation metrics of tasks in CVLUE. R@k denotes the recall in the top k predictions, Acc stands for accuracy, and IoU stands for intersection over union.

Semantic Fields	Categories
Animal	panda, *cow*, fish, *dog*, *horse*, chicken, *mouse*, *bird*, *human*, *cat*
Food	**hot pot**, rice, dumpling, noodles, **stuffed bun**
Beverages	**bubble tea**, coke, milk, tea, porridge, alcohol
Clothing	**Hanfu**, **Tang suit**, **cheongsam**, suit, T-shirt
Plant	willow, ginkgo, Chinese parasol, birch, pine, chrysanthemum, peony, orchid, lotus, lily
Fruit	lychee, hawthorn, *apple*, cantaloupe, longan
Vegetable	bok choy, potato, Chinese cabbage, carrot, *cauliflower*
Agriculture Tool	hoe, plow, harrow, sickle, **carrying pole**
Furniture	*spoon*, *bowl*, cutting board, chopsticks, wok, fan, **Chinese cleaver**, **wok spatula**
Sport	*TV*, table, *chair*, *refrigerator*, cooking stove
Celebrations	Ping-Pong, basketball, swimming, football, running
Education	**lion dance**, **dragon boat**, national flag, **mooncake**, couplet, lantern
Instruments	pencil, blackboard, **Chinese brush**, chalk, ballpoint, *scissors*
Arts	**Chinese zither**, **erhu**, **suona**, drums, **pipa** **brush calligraphy**, **Chinese shadow play**, **paper cutting**, **Terracotta Army**, **ding**, ceramics

Table 3: Object categories in CVLUE, where the 15 categories overlapping with MS-COCO are shown in blue italic font, while the 22 categories not in WordNet are shown in red bold font.

Selection of Object Categories

Our selection of object categories aimed for a representative set in Chinese daily life that reflected the unique characteristics of Chinese culture. The selection process was inspired by the Chinese part of MaRVL (Liu et al. 2021), where five native speakers provided 5-10 concepts for 18 semantic fields, ensuring they are commonly seen and representative. However, since CVLUE is specifically for Chinese, MaRVL's categories are not directly applicable.

Therefore, we first removed categories not strongly related to specific objects with clear boundaries (e.g., Taoism). We also replaced some categories with more concrete categories that have clearer boundaries (e.g., replacing the Mid-Autumn Festival with moon cake). Then, we merged some categories to make sure that all categories occurred frequently enough so that we could collect enough images for each of them (e.g., merging all types of birds into one bird category). Besides, we added some categories representative of Chinese culture (e.g., stuffed buns, fans).[2]

Eventually, we selected 92 object categories from 15 semantic fields listed in Table 3. The 15 categories overlapping with MS-COCO (e.g., human, dog), shown in blue italic font, can be regarded as having the weakest association with Chinese culture. The 22 categories not in English WordNet (Miller 1992) (e.g., guzheng, suona), shown in red bold font, are considered to be culturally closest to Chinese.

[2]See the Appendix for the original categories in MaRVL.

The remaining categories have a moderate association.

Task Selection

As introduced in the related work section, there are currently a wide variety of VL tasks. Due to budgetary constraints, we focused on the following four pivotal and representative VL tasks for our dataset:[3]

Image-Text Retrieval: This task includes text retrieval, where given an image, the task is to retrieve the corresponding text, and image retrieval, where given a text, the task is to retrieve the corresponding image. It evaluates VLMs' ability to align vision and language representations.

Visual Question Answering: Given an image and a natural language question, the model must generate a correct answer. It assesses VLMs' detailed visual understanding and reasoning skills.

Visual Grounding: Given an image and a referring expression, the model must locate the specified object. This task measures VLMs' ability to understand and identify objects in images.

Visual Dialog: Given an image, a dialogue history, and a question about the image, the model must answer accurately. This task evaluates VLMs' overall intelligence, including visual understanding, memory, and language generation.

Image Collection

After obtaining the list of object categories, our next goal was to collect appropriate images for each of them. To meet the requirements of different types of tasks in our dataset, we collected two subsets of images for each category. Subset A consists of images *containing at least 2 objects of the same category* and is used for the VQA and VG tasks. Subset B consists of images *containing 3-5 objects of different object categories* and is used for the VD task. The ITR task is annotated on both subsets. All the collected images must be (1) real photos with no watermark; (2) non-iconic images with more than 2 objects; (3) commonly seen or representative in Chinese culture. We used the Baidu data crowdsourcing platform for image collection, where Chinese native speakers gathered images from the Chinese Internet according to the guidelines mentioned above, ensuring all images complied with our required open-source license. Images were first inspected by the platform's quality control inspectors and then reviewed by the authors. Any images not meeting the guidelines were discarded. During annotation, annotators were also allowed to discard non-compliant images.

Quality Control

We used a two-step process for selecting and training annotators to ensure annotation quality. First, candidates received annotation guidelines and annotated five images to assess their general capability. Qualified candidates were then grouped by task based on their performance. Second, each group annotated 50 randomly sampled images, guided one-on-one by senior annotators until they fully understood the guidelines and achieved 100% accuracy on these images.

[3]See the Appendix for the detailed annotation process.

Annotators who completed the training began annotating tasks batched into packages. They could not proceed to the next package until finishing the current one. Each package was *self-checked, reviewed by a senior inspector, and eventually inspected by four co-authors familiar with the guidelines*. The final inspection sampled 10%-25% of each package, requiring over 97% accuracy to pass. Otherwise, the package was returned for correction. The IC, VQA, VG, and VD tasks involved 41, 108, 44, and 26 annotators and 10, 12, 8, and 13 senior inspectors, respectively. The project took six months and cost approximately RMB 550,000.

Data Characteristics

In this section, we analyse the annotated data to show their characteristics.

Images and Objects

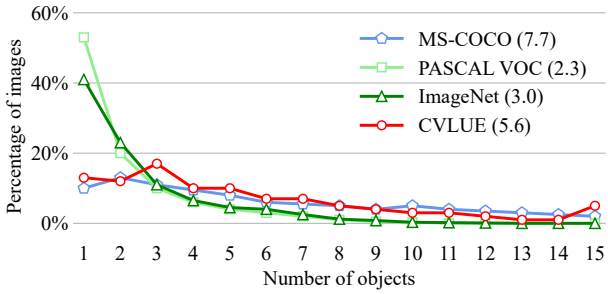


Figure 2: Number of annotated objects per image for CVLUE, MS-COCO, ImageNet Detection and PASCAL VOC (average numbers are shown in parentheses).

We first count the object-related statistics to show the properties of the source images in CVLUE. The number of objects per category for all 92 categories is shown in the Appendix. We compare CVLUE with several popular datasets, including MS-COCO (Lin et al. 2014), ImageNet (Deng et al. 2009) and PASCAL VOC (Everingham et al. 2010). These datasets have different purposes: MS-COCO for detecting and segmenting objects in context, ImageNet for capturing object categories, and PASCAL VOC for detecting objects in natural images. CVLUE, however, is specifically designed to evaluate VLMs comprehensively in Chinese VLU. The numbers of annotated objects per image are shown in Figure 2. Our dataset averages 5.6 annotated objects per image, compared to less than 3 for ImageNet and PASCAL VOC.

Image Text Retrieval

For the ITR task, we compare CVLUE with several popular Chinese datasets constructed via text translation (Flickr30K) or re-annotation (Flickr8K and COCO-CN). These datasets are all built on top of Western culture-biased images from existing English VL datasets. The caption length distribution is shown in Figure 3. Our dataset's average caption length is 19.2, which is higher than that of the other three datasets. It

Tasks	Dataset	Fine-tuning		Zero-shot		
		CCLM 522M	X^{2}VLM 422M	QwenVL 7B	QwenVL-Chat 7B	mPLUG-Owl2 7B
TR	COCO (5K)	77.7	**80.1**	-	-	-
	CVLUE	49.9	**54.8**	-	-	-
IR	COCO (5K)	60.5	**63.8**	-	-	-
	CVLUE	32.0	**36.6**	-	-	-
VQA	VQA-v2 (test-std)	63.7	75.5	78.0	67.9	**79.2**
	CVLUE	**58.5**	53.0	29.9	39.8	20.4
VG	RefCOCOg	70.4	79.9	78.0	**80.1**	-
	CVLUE	39.1	**48.8**	36.8	40.4	-
VD	Visdial 1.0	**42.4**	41.5	36.0	37.5	37.2
	CVLUE	**32.2**	27.6	24.8	26.5	25.8

Table 4: Results of baseline VLMs. We report R@1 for the TR, IR and VD tasks, accuracy for the VQA task and IoU for the VG task. Number of parameters for each model is listed below its name. The bold font indicates the best result for each task.

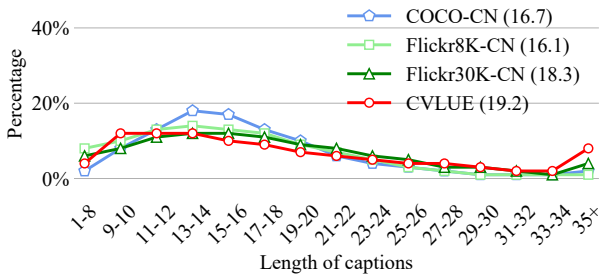


Figure 3: The caption length distribution of CVLUE, COCO-CN, Flickr8K-CN and Flickr30K-CN (average caption lengths are shown in parentheses).

is worth noting that the caption lengths in CVLUE are distributed more evenly than the other three datasets. This indicates that our dataset comprises both simple captions and complicated ones.

Visual Grounding

To the best of our knowledge, there has not been any other Chinese VG dataset. To illustrate our dataset's properties, here we provide a rough comparison between the VG dataset in CVLUE and a popular English VG dataset RefCOCOg (Mao et al. 2016). Overall, the average number of referring expressions per image is 3.38 for our VG dataset and 3.91 for RefCOCOg. This is because multiple expressions for a single object are allowed in RefCOCOg but disallowed in our dataset. The average number of objects described per image in our dataset and in RefCOCOg is 3.38 and 1.93, respectively, meaning that more objects are described in our dataset. Besides, the average expression lengths are 11.9 characters for our dataset and 8.3 words for RefCOCOg.

Experiments

Experimental Setups and Baselines

We used CVLUE and some of its English counterparts to evaluate the performance of several popular multilingual VLMs in VLU. The English VL datasets include COCO (5K) (Lin et al. 2014), VQA-v2 (Goyal et al. 2017), RefCOCOg (Mao et al. 2016) and Visdial 1.0 (Das et al. 2017).[4]

Due to limited budget and computational resources, we couldn't test all VLMs, especially with their rapid increase in variety. Therefore, we selected some popular and representative models, covering both fine-tuning and zero-shot settings, and left the analysis of more VLMs for future work. Models under the fine-tuning setting include:

CCLM (Zeng et al. 2023), a multilingual VLM where the cross-lingual and cross-modal objectives are jointly learned.

X^{2}VLM (Zeng et al. 2022), a multilingual VLM where the multi-grained vision language alignments are learned in a unified framework.

Models under the zero-shot setting include:

Qwen-VL (Bai et al. 2023), a large-scale VLM pre-trained on 7 VL tasks simultaneously, can handle the grounding task.

Qwen-VL-Chat, the Qwen-VL model fine-tuned through instruction tuning with the instruction following and dialogue capabilities enhanced.

mPLUG-Owl2 (Ye et al. 2023), a large-scale VLM that incorporates shared functional modules to facilitate modality collaboration.

We couldn't afford to tune hyper-parameters for each baseline model, so we used default ones for them all. Please refer to the Appendix for prompts used in the zero-shot setting and detailed fine-tuning setups. For the VD task, we collected 100 candidate answers (including correct, plausible, popular and random ones) for each question following the procedure proposed by Das et al. (2017).

[4]We use the default splits for these datasets.

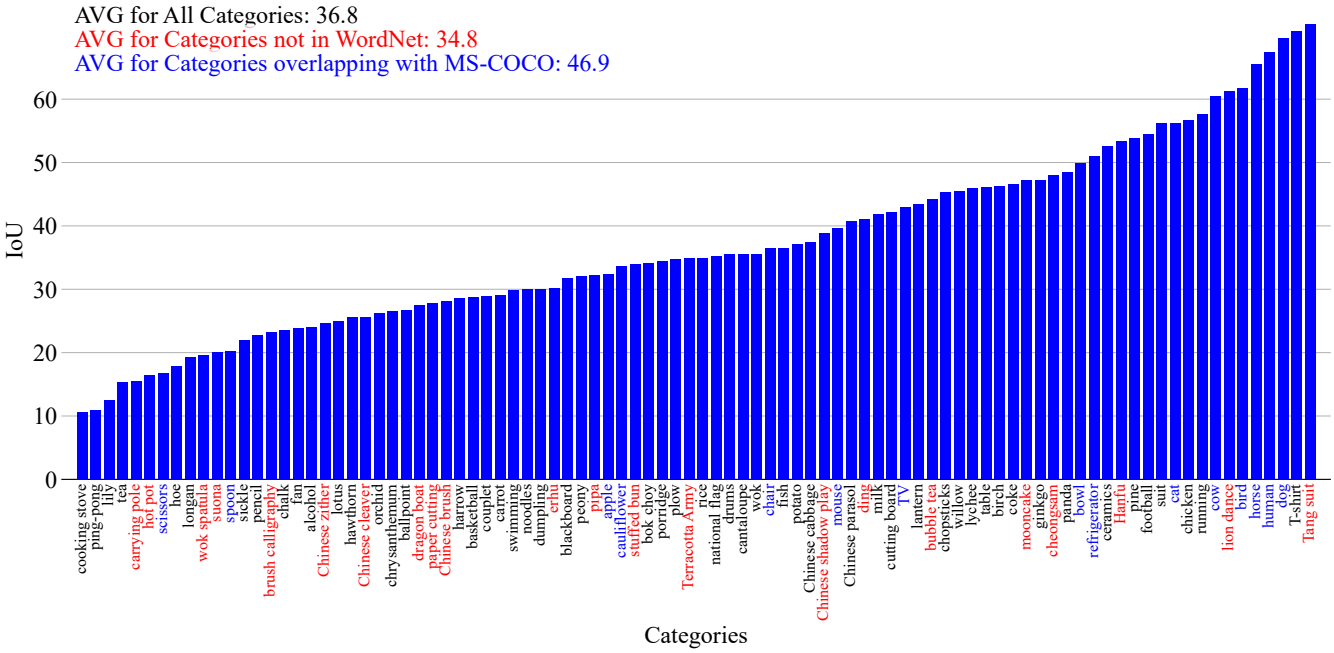


Figure 4: Results of QwenVL model on the CVLUE VG task, displayed by image category.

Results

The results of the baseline models on CVLUE are presented in Table 4.[5] All models under the zero-shot setting do not support the ITR task. Additionally, mPLUG-Owl2 does not support the VG task. Hence, these results are not reported.

The three large-scale VLMs under the zero-shot setting yield strong performance on the English datasets they are evaluated on, and some of their results are even higher than those of the two models under the fine-tuning setting. This could be attributed to their larger model capacity and the fact that they have been pre-trained on various VL tasks. On the other hand, all five models' performance on CVLUE is much lower than that on the English VL dataset. This aligns with the results observed on CCBench discussed in the related work section. Such a substantial performance gap between English and Chinese VL datasets indicates that the VLU capability of existing multilingual VLMs (under both zero-shot and fine-tuning settings) in Chinese severely lags behind that in English. We conduct case studies on model predictions for VQA and VG tasks under zero-shot and fine-tuning settings. Our findings indicate that errors in zero-shot settings often stem from a lack of Chinese cultural knowledge. In contrast, the fine-tuned CCLM and X2VLM models demonstrate improved Chinese VLU capabilities, likely due to the enhancement of Chinese cultural knowledge during the fine-tuning process. Relevant case studies are provided in the Appendix.

Besides, we find that on CVLUE, zero-shot models, despite having more parameters, often perform worse than fine-tuned models. Conversely, on English VL tasks, zero-shot models sometimes outperform fine-tuned ones. We believe this is because these zero-shot models inherently possess more Western cultural knowledge than Chinese cultural knowledge, and their larger parameter scale gives them an edge in English tasks.

[5]See the Appendix for full results containing R@5 and R@10 for the TR, IR and VD tasks.

Analysis

Results by Category

To comprehensively investigate existing VLMs' VLU capabilities regarding Chinese culture, the first question to address is *whether existing VLMs truly exhibit a significant performance difference between categories that are closely related to Chinese culture and those that are less related.*

Our dataset provides category information for each image, allowing for a fine-grained analysis of results across different categories. This facilitates the precise identification of the specific image categories in which VLMs exhibit deficiencies in their VLU abilities. As discussed in the 'selection of object categories' section, the 92 categories in CVLUE can be roughly divided into three groups: 1) categories culturally closest to Chinese (i.e., those not in WordNet), 2) categories with the weakest association with Chinese culture (i.e., those overlapping with MS-COCO) and 3) categories with moderate association (i.e., the remaining ones). To answer the question, we analyse the models' results across different categories.

Figure 4 shows the performance of the QwenVL model on the VG task, displayed by category. The results for categories closely related to Chinese culture are generally lower, with an average score of 34.8, while the results for categories overlapping with MS-COCO are generally higher, with an

average score of 46.9.[6] This performance gap highlights a clear deficiency in existing VLMs' VLU capabilities regarding Chinese culture.

Results on Translated English Test Sets

Given that a majority of existing VL data used for pre-training focus on English with predominantly Western-centric images, the next question is *whether the knowledge required to address tasks closely related to Chinese culture is present in the English part of existing VLMs*.

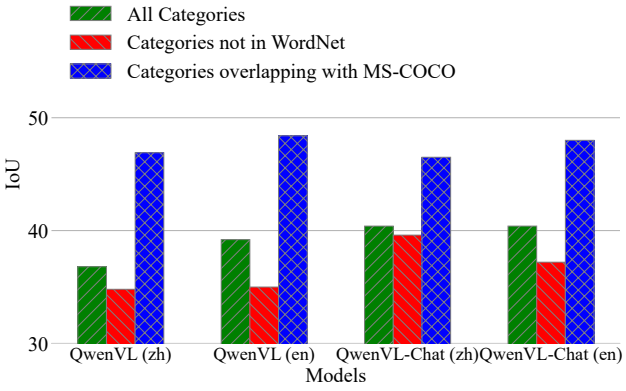


Figure 5: Category group results of QwenVL and QwenVL-Chat on the original Chinese (zh) and translated English (en) CVLUE VG test set.

To address this question, we translate the VG test set into English, and then compare QwenVL and QwenVL-Chat predictions with their results on the original Chinese test set. According to Figure 5, for the same model, when the test set is translated from Chinese to English, performance on categories closely related to Chinese culture (not in WordNet) often remains unchanged or declines, while performance on categories less related to Chinese culture (overlapping with MS-COCO) significantly improves.[7] This indicates that in these VLMs, the English part typically contains more knowledge of categories less related to Chinese culture but, like the Chinese part, lacks knowledge of categories closely related to Chinese culture.

Zero-Shot vs. Fine-Tuning

Due to the lack of knowledge required to address tasks closely related to Chinese culture in both the Chinese and English parts of existing VLMs, the final question becomes *how to effectively enhance the knowledge of Chinese culture in these VLMs*.

In this section, we compare the performance of models under the zero-shot and the fine-tuning settings. According to the results on the CVLUE VG task in Figure 6, Chinese culture-related categories perform significantly lower than average on zero-shot models but higher than average on fine-tuned models.[8] This indicates that fine-tuning with CVLUE's Chinese cultural VL data benefits categories strongly related to Chinese culture more. Overall, fine-tuning on Chinese cultural VL data is an effective way to enhance the VLM's VLU capabilities regarding Chinese culture.

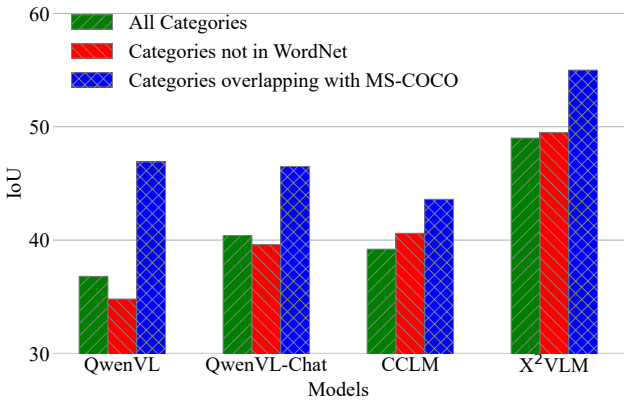


Figure 6: Category group results on CVLUE VG task.

Conclusion

In this paper, we present CVLUE, a vision-language understanding benchmark dataset specifically designed for the comprehensive evaluation of VLMs in Chinese VLU. Images used in the dataset are newly collected by Chinese native speakers with explicit constraints ensuring that they are representative of Chinese culture and thus avoid the cultural bias caused by exploiting Western-centric images from existing English VL datasets. Four distinct and representative VL tasks are included in CVLUE for the multi-aspect evaluation of VLMs in Chinese culture. Using CVLUE and some English VL datasets, we reveal a noticeable gap between the performance of several strong multilingual VLMs on English and Chinese VLU. Our in-depth category-level analysis reveals a lack of Chinese culture-related knowledge in existing VLMs and shows that fine-tuning on Chinese culture-related VL datasets can effectively enhance VLMs' VLU capabilities regarding Chinese culture. We believe that CVLUE is a solid step towards a fair and convenient platform for the comparison of VLMs in Chinese culture and can eventually facilitate the development of Chinese vision-language pre-training.

Ethical Statement

Images in our dataset were collected on the Baidu data crowdsourcing platform from the Chinese Internet using a purely manual approach, with a requirement that all submitted images comply with the CC BY-NC-ND 4.0 license. For categories where it was unable to collect enough images that comply with this license, we commissioned the platform to obtain the remaining images from third-party image suppliers. Sensitive information in the images (e.g., human faces) has been obscured to prevent potential misuse of the dataset. We used the Baidu data crowdsourcing platform for annotation. All the annotators have given informed consent and

[6] Similar pattern observed on other tasks in the Appendix.
[7] Similar pattern observed on VQA in the Appendix.
[8] Similar pattern observed on VQA in the Appendix.

have been fairly compensated during the image collection and annotation process. The proposed dataset will be made publicly available for research purposes (under the CC BY-NC-ND 4.0 license) after the paper gets accepted.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (No. 2022YFB4500300) and the Key Research Project of Zhejiang Lab (No. 2024SSYS0005).

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proc. of ICCV*, 2425–2433.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*.

Burda-Lassen, O.; Chadha, A.; Goswami, S.; and Jain, V. 2024. How Culturally Aware are Vision-Language Models? *CoRR*.

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR*.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Proc. of ECCV*, 104–120.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proc. of ICML*, 1931–1942.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *Proc. of CVPR*, 1080–1089.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*, 248–255.

DeVries, T.; Misra, I.; Wang, C.; and van der Maaten, L. 2019. Does Object Recognition Work for Everyone? In *Proc. of CVPR*, 52–59.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 303–338.

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *CoRR*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proc. of CVPR*, 6325–6334.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; Xu, C.; and Xu, H. 2022. Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark. In *Proc. of NeurIPS*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. L. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proc. of EMNLP*, 787–798.

Lan, W.; Li, X.; and Dong, J. 2017. Fluency-Guided Cross-Lingual Image Captioning. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 1549–1557.

Li, H.; Jiang, L.; Huang, J. D.; Kim, H.; Santy, S.; Sorensen, T.; Lin, B. Y.; Dziri, N.; Ren, X.; and Choi, Y. 2024. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. *CoRR*.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Proc. of NeurIPS*, 9694–9705.

Li, X.; Lan, W.; Dong, J.; and Liu, H. 2016. Adding Chinese Captions to Images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*, 271–275.

Li, X.; Xu, C.; Wang, X.; Lan, W.; Jia, Z.; Yang, G.; and Xu, J. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *IEEE Trans. Multim.*, 2347–2360.

Lin, J.; Men, R.; Yang, A.; Zhou, C.; Ding, M.; Zhang, Y.; Wang, P.; Wang, A.; Jiang, L.; Jia, X.; Zhang, J.; Zhang, J.; Zou, X.; Li, Z.; Deng, X.; Liu, J.; Xue, J.; Zhou, H.; Ma, J.; Yu, J.; Li, Y.; Lin, W.; Zhou, J.; Tang, J.; and Yang, H. 2021. M6: A Chinese Multimodal Pretrainer. *CoRR*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*, 740–755.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proc. of EMNLP*, 10467–10485.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; Chen, K.; and Lin, D. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *CoRR*.

Liu, Y.; Zhu, G.; Zhu, B.; Song, Q.; Ge, G.; Chen, H.; Qiao, G.; Peng, R.; Wu, L.; and Wang, J. 2022. TaiSu: A 166M Large-scale High-Quality Dataset for Chinese Vision-Language Pre-training. In *Proc. of NeurIPS*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proc. of NeurIPS*, 13–23.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *Proc. of CVPR*, 11–20.

Miller, G. A. 1992. WordNet: A Lexical Database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Parcalabescu, L.; Cafagna, M.; Muradjan, L.; Frank, A.; Calixto, I.; and Gatt, A. 2022. VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena. In *Proc. of ACL*, 8253–8280.

Qi, L.; Lv, S.; Li, H.; Liu, J.; Zhang, Y.; She, Q.; Wu, H.; Wang, H.; and Liu, T. 2022. DuReader$_{vis}$: A Chinese Dataset for Open-domain Document Visual Question Answering. In *Proc. of ACL Findings*, 1338–1351.

Rajendran, J.; Khapra, M. M.; Chandar, S.; and Ravindran, B. 2016. Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning. In *Proc. of NAACL*, 171–181.

Rao, A.; Yerukola, A.; Shah, V.; Reinecke, K.; and Sap, M. 2024. NORMAD: A Benchmark for Measuring the Cultural Adaptability of Large Language Models. *CoRR*.

Romero, D.; Lyu, C.; Wibowo, H. A.; Lynn, T.; Hamed, I.; Kishore, A. N.; Mandal, A.; Dragonetti, A.; Abzaliev, A.; Tonja, A. L.; Balcha, B. F.; Whitehouse, C.; Salamea, C.; Velasco, D. J.; Adelani, D. I.; Meur, D. L.; Villa-Cueva, E.; Koto, F.; Farooqui, F.; Belcavello, F.; Batnasan, G.; Vallejo, G.; Caulfield, G.; Ivetta, G.; Song, H.; Ademtew, H. B.; Maina, H.; Lovenia, H.; Azime, I. A.; Cruz, J. C. B.; Gala, J.; Geng, J.; Ortiz-Barajas, J.-G.; Baek, J.; Dunstan, J.; Alemany, L. A.; Nagasinghe, K. R. Y.; Benotti, L.; D'Haro, L. F.; Viridiano, M.; Estecha-Garitagoitia, M.; Cabrera, M. C. B.; Rodríguez-Cantelar, M.; Jouitteau, M.; Mihaylov, M.; Imam, M. F. M.; Adilazuarda, M. F.; Gochoo, M.; Otgonbold, M.-E.; Etori, N.; Niyomugisha, O.; Silva, P. M.; Chitale, P.; Dabre, R.; Chevi, R.; Zhang, R.; Diandaru, R.; Cahyawijaya, S.; Góngora, S.; Jeong, S.; Purkayastha, S.; Kuribayashi, T.; Jayakumar, T.; Torrent, T. T.; Ehsan, T.; Araujo, V.; Kementchedjhieva, Y.; Burzo, Z.; Lim, Z. W.; Yong, Z. X.; Ignat, O.; Nwatu, J.; Mihalcea, R.; Solorio, T.; and Aji, A. F. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *CoRR*.

Srinivasan, T.; Chang, T.; Alva, L. L. P.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. CLiMB: A Continual Learning Benchmark for Vision-and-Language Tasks. In *Proc. of NeurIPS*.

Stock, P.; and Cissé, M. 2018. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In *Proc. of ECCV*, 504–519.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proc. of ACL*, 6418–6428.

Wang, B.; Lv, F.; Yao, T.; Ma, J.; Luo, Y.; and Liang, H. 2022. ChiQA: A Large Scale Image-based Real-World Question Answering Dataset for Multi-Modal Understanding. In *Proc. of CIKM*, 1996–2006.

Wang, W.; Jiao, W.; Huang, J.; Dai, R.; Huang, J.; Tu, Z.; and Lyu, M. R. 2023. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. *CoRR*.

Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.; Wang, Y.; and Wang, Y. 2017. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. *CoRR*.

Xie, C.; Cai, H.; Song, J.; Li, J.; Kong, F.; Wu, X.; Morimitsu, H.; Yao, L.; Wang, D.; Leng, D.; Ji, X.; and Deng, Y. 2022. Zero and R2D2: A Large-scale Chinese Cross-modal Benchmark and A Vision-Language Framework. *CoRR*.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. *CoRR*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *CoRR*.

Yoshikawa, Y.; Shigeto, Y.; and Takeuchi, A. 2017. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In *Proc. of ACL*, 417–421.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 67–78.

Zeng, Y.; Zhang, X.; Li, H.; Wang, J.; Zhang, J.; and Zhou, W. 2022. X^2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks. *CoRR*.

Zeng, Y.; Zhou, W.; Luo, A.; Cheng, Z.; and Zhang, X. 2023. Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training. In *Proc. of ACL*, 5731–5746.

Zhan, X.; Wu, Y.; Dong, X.; Wei, Y.; Lu, M.; Zhang, Y.; Xu, H.; and Liang, X. 2021. Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining. In *Proc. of ICCV*, 11762–11771.

Zhao, W.; Mondal, D.; Tandon, N.; Dillion, D.; Gray, K.; and Gu, Y. 2024. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In *Proc. of COLING*, 17696–17706.

Zheng, K.; Chen, X.; Jenkins, O. C.; and Wang, X. 2022. VLMbench: A Compositional Benchmark for Vision-and-Language Manipulation. In *Proc. of NeurIPS*.

Zhou, W.; Zeng, Y.; Diao, S.; and Zhang, X. 2022. VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models. *CoRR*, abs/2205.15237.