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

Anonymous ACL submission

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. [1]

1 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 is a small body of work 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

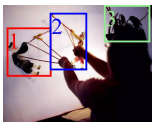
[1] Our benchmark and the evaluation codes will be released after the paper gets accepted.

1. 桌上放着两口火锅和一些火锅食材
(Two hotpots and some hotpot ingredients are on the table)
2.火锅周围的食物和餐具摆满了整个桌面
(Food and utensils around the hotpots cover the table)
3.煮有食物的两个火锅正在沸腾，锅里煮着肉丸和豆腐
(The two hotpots are boiling, with meatballs and tofu inside)
4.食材中间放着两口煮着蔬菜和肉类的火锅，两口火锅均呈白色
(Among the food, there are two white hotpots cooking vegetables and meat)
5.两个烧开的火锅周围摆放着一些肉类和蔬菜，它们的边上放着碗筷和玻璃杯
(Around the boiling hotpots, there are meat, vegetables, bowls, chopsticks, and glasses)

(a) Image Text Retrieval

Q: 人群的两边是什么农作物？
(What crops are on both sides of the crowd?)
A: 稻谷
(Rice)
Q: 舞龙的人数量多还是舞狮的人数量多？
(Are there more people in the dancing dragon or in the dancing lion?)
A: 舞龙的
(Dancing dragon)
Q: 舞龙的数量比舞狮的数量少多少？
(How many fewer dancing dragons are there compared to dancing lions?)
A: 1

(b) Visual Question Answering



1.人左手操控的皮影
(The shadow puppet manipulated by the person's left hand)
2.人右手操控的皮影
(The shadow puppet manipulated by the person's right hand)
3.没有人操控的皮影
(The shadow puppet not manipulated by anyone)

(c) Visual Grounding

Caption: 龙舟的船头坐着一个戴墨镜的男子在指挥
(A man wearing sunglasses is sitting at the front of a dragon boat, directing)
Q1: 龙舟上大约有多少人？
(How many people are on the dragon boat?)
A1: 龙舟上大约有二十来个人
(There are about twenty people on the dragon boat)
……
Q10: 龙舟是什么颜色的？
(What color is the dragon boat?)
A10: 船身是蓝色的，船头船尾是黄红色的
(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 for the four tasks in CVLUE.

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 dataset 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 et al., 2017), were 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), were constructed by re-annotating images from English VL datasets in Chinese. Recently, several new datasets have been presented, whose images were collected from image search engines with Chinese queries. 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, whose 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.[2] 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[3], 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 the images and annotation for the four tasks are shown in Figure 1. See Appendix A.4 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.

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

[3]https://tianchi.aliyun.com/muge

2 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) and 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 et al., 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 et al., 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, given the situation of existing Chinese VL datasets, building a benchmark specifically for Chinese is 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 MMBench, which is first annotated in English and then translated to Chinese using GPT-4. Interestingly, they also released CCBench[4], 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.

3 CVLUE

Our dataset 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 3. In this section, we describe the procedure we devised for image collection and dataset annotation.

3.1 Selection of Object Categories

We first explain the selection of object categories, which must form a representative set of categories in Chinese daily life and reflect the unique characteristics of Chinese culture. The selection process for our dataset was inspired by the Chinese part of MaRVL (Liu et al., 2021), where five native speakers provided 5-10 specific concepts for 18 semantic fields, ensuring they are commonly seen, representative, physical, and concrete. 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 cat-

[4] https://github.com/open-compass/MMBench

Semantic Fields	Categories
Animal	大熊猫 (panda), 牛 *(cow)*, 鱼 (fish), 狗 *(dog)*, 马 *(horse)*, 鸡 (chicken), 鼠 *(mouse)*, 鸟 *(bird)*, 人 *(human)*, 猫 *(cat)*
Food	火锅 **(hot pot)**, 米饭 (rice), 饺子 (dumpling), 面条 (noodles), 包子 **(stuffed bun)**
Beverages	奶茶 **(bubble tea)**, 可乐 (coke), 牛奶 (milk), 茶 (tea), 粥 (porridge), 酒 (alcohol)
Clothing	汉服 **(Hanfu)**, 唐装 **(Tang suit)**, 旗袍 **(cheongsam)**, 西装 (suit), T恤 (T-shirt)
Plant	柳树 (willow), 银杏 (ginkgo), 梧桐 (Chinese parasol), 白桦 (birch), 松树 (pine), 菊花 (chrysanthemum), 牡丹 (peony), 兰花 (orchid), 莲 (lotus), 百合 (lily)
Fruit	荔枝 (lychee), 山楂 (hawthorn), 苹果 *(apple)*, 哈密瓜 (cantaloupe), 龙眼 (longan)
Vegetable	小白菜 (bok choy), 马铃薯 (potato), 大白菜 (Chinese cabbage), 胡萝卜 (carrot), 花椰菜 *(cauliflower)*
Agriculture	锄头 (hoe), 犁 (plow), 耙 (harrow), 镰刀 (sickle), 担杖 **(carrying pole)**
Tool	汤勺 *(spoon)*, 碗 *(bowl)*, 砧板 (cutting board), 筷子 (chopsticks), 炒锅 (wok), 扇子 (fan), 菜刀 **(Chinese cleaver)**, 锅铲 **(wok spatula)**
Furniture	电视 *(TV)*, 桌子 (table), 椅子 *(chair)*, 冰箱 *(refrigerator)*, 灶台 (cooking stove)
Sport	乒乓球 (Ping-Pong), 篮球 (basketball), 游泳 (swimming), 足球 (football), 跑步 (running)
Celebrations	舞狮 **(lion dance)**, 龙舟 **(dragon boat)**, 国旗 (national flag), 月饼 **(mooncake)**, 春联 (couplet), 花灯 (lantern)
Education	铅笔 (pencil), 黑板 (blackboard), 毛笔 **(Chinese brush)**, 粉笔 (chalk), 原子笔 (ballpoint), 剪刀 *(scissors)*
Instruments	古筝 **(Chinese zither)**, 二胡 **(erhu)**, 唢呐 **(suona)**, 鼓 (drums), 琵琶 **(pipa)**
Arts	毛笔书法 **(brush calligraphy)**, 皮影 **(Chinese shadow play)**, 剪纸 **(paper cutting)**, 兵马俑 **(Terracotta Army)**, 鼎 **(ding)**, 陶瓷 (ceramics)

Table 2: 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.

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 3: 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.

egories with more concrete categories that have clearer boundaries (e.g., replacing the Dragon Boat Festival with dragon boat, 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).

Eventually, we selected 92 object categories from 15 semantic fields listed in Table 2. 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. The remaining categories have a moderate association.

3.2 Task Selection

As introduced in section 2, 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:[5]

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.

3.3 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 differ-

[5]See Appendix A.2 for the detailed annotation process.

ent types of tasks in our dataset, we collect 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.[6] Subset B consists of images *containing 3-5 objects of different object categories* and is used for the VD task.[7] The image captioning 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. The images were collected from the Chinese Internet and inspected by four co-authors who are well aware of the image collection guidelines.

3.4 Quality Control

To ensure annotation quality, we use a two-step process for selecting and training annotators. First, candidates receive annotation guidelines and annotate five randomly sampled images to assess their general capability. Qualified candidates are then grouped by task based on their performance. Second, each group annotates 50 randomly sampled images, guided one-on-one by senior annotators until they fully understand the guidelines and achieve 100% accuracy on these 50 images.

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.

4 Data Characteristics

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

4.1 Images and Objects

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 Appendix A.1. We compare

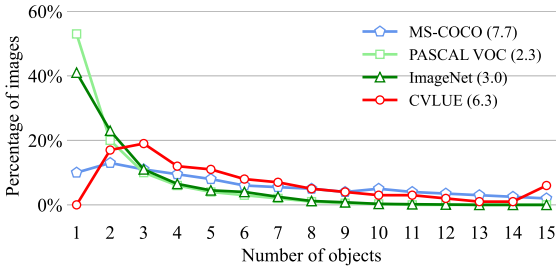


Figure 2: Number of annotated objects per image for CVLUE, MS-COCO, ImageNet Detection and PASCAL VOC (average numbers are shown in parentheses).

CVLUE with several popular datasets, including MS-COCO (Lin et al., 2014), ImageNet[8] (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. Our dataset averages 6.3 annotated objects per image, compared to less than 3 for ImageNet and PASCAL VOC. Notably, no CVLUE images contain only one object due to subset A's requirement of at least two objects of the same category per image. The numbers of annotated objects per image are shown in Figure 2. Our dataset averages 6.3 annotated objects per image, compared to less than 3 for ImageNet and PASCAL VOC. Notably, no CVLUE images contain only one object due to subset A's requirement of at least two objects of the same category per image.

4.2 Image Text Retrieval

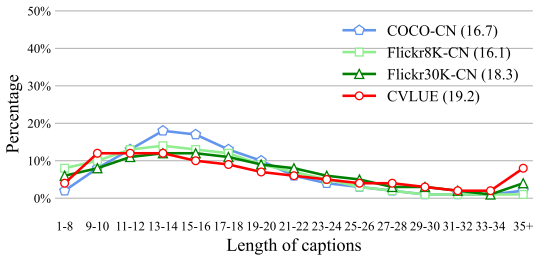


Figure 3: The caption length distribution of CVLUE, COCO-CN, Flickr8K-CN and Flickr30K-CN (average caption lengths are shown in parentheses).

[6]This constraint ensures VG is challenging enough.

[7]This constraint improves the richness of dialogues in VD.

[8]We use the object detection validation set since the training data only has a single object labelled.

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. For each compared model, we also report the number of parameters.

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 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.

4.3 Visual Grounding

To the best of our knowledge, there has not been any other Chinese VG dataset. To illustrate the property of the proposed dataset, 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.

5 Experiments

5.1 Experimental Setups and Baselines

We use CVLUE and some of its counterparts in English 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).[9]

We use two experimental settings, namely the fine-tuning one and the zero-shot one. 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 Appendix A.7 and A.5 for prompts used in the zero-shot setting and detailed fine-tuning setups. For the VD task, we collect 100 candidate answers (including correct, plausible, popular and random ones) for each question following the procedure proposed by Das et al. (2017).

5.2 Results

The results of the baseline models on CVLUE are presented in Table 4.[10] All models under the zero-

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

[10] See Appendix A.6 for full results containing R@5 and R@10 for the TR, IR and VD tasks.

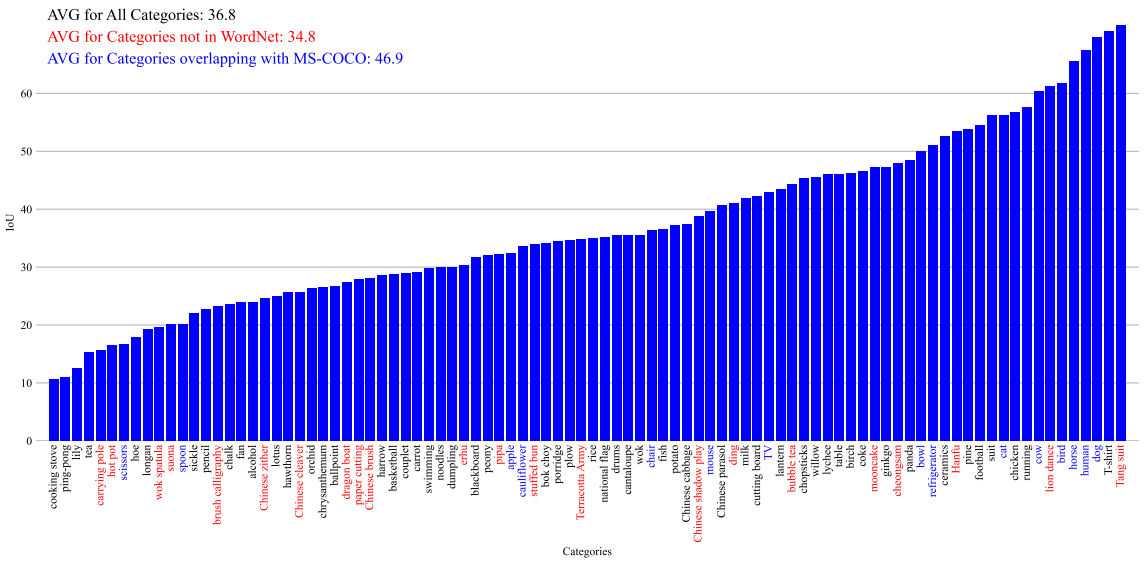


Figure 4: Results of QwenVL model on the CVLUE VG task, displayed by image category.

shot setting do not support the ITR task. Additionally, mPLUG-Owl2 does not support the VG task either. 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 section 2. 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. 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 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.

6 Analysis

6.1 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 section 3.1, 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 analyze 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.[11] This performance gap highlights a clear deficiency in existing VLMs' VLU capabilities regarding Chinese culture.

6.2 Results on Translated English Test Sets

Given that a majority of existing VL data used for pre-training focus on English with predomi-

[11] Similar pattern observed on other tasks in Appendix A.10.

nantly 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*.

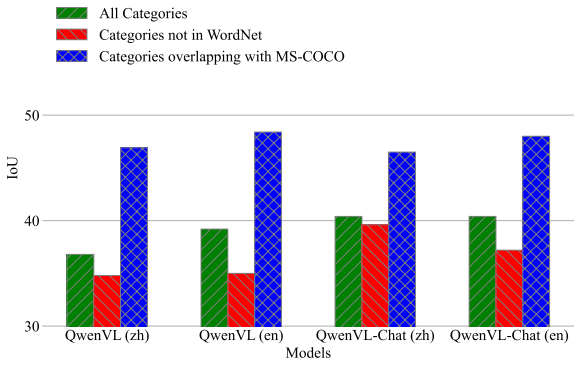


Figure 5: Category group results of QwenVL and QwenVL-Chat on the original Chinese (zh) and translated English (en) CVLUE VG test set.

To address this question, we use GPT-4 to translate the VG test set into English, 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.[12] 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.

6.3 Zero-Shot vs. Fine-Tuninig

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.[13] This indicates that fine-tuning with CVLUE's Chinese cultural VL data benefits

[12]Similar pattern observed on VQA in Appendix A.8.
[13]Similar pattern observed on VQA in Appendix A.9.

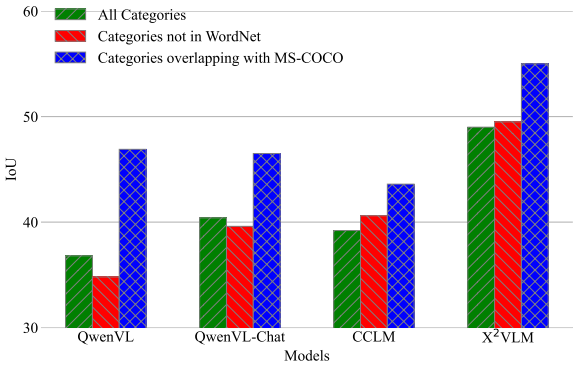


Figure 6: Category group results on CVLUE VG task.

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.

7 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 were 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 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.

8 Ethical Considerations

Images used in our benchmark are collected from the Chinese Internet. 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 image collection and annotation. All the annotators have

given informed consent and 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-ND license) after the paper gets accepted.

9 Limitations

Due to limited computational resources, we were unable to test all VLMs or fine-tune larger VLMs on the proposed dataset. Therefore, we selected some popular and representative models and conducted experiments under both fine-tuning and zero-shot settings. Additionally, we couldn't afford to tune hyperparameters for each model, so we used the same default settings for all. Consequently, the reported results may not reflect the models' full potential. However, we believe that the current experimental setup is sufficient to highlight the significant performance gap between English and Chinese VL datasets for these strong and popular VLMs. Furthermore, the in-depth category-level analysis demonstrates that existing VLMs lack knowledge related to Chinese culture, validating the usefulness of CVLUE for comprehensive and fine-grained evaluation of VLMs in Chinese VLU.

Additionally, some may argue that the four tasks included in CVLUE are too few for a comprehensive evaluation of VLMs. However, due to budgetary constraints and to ensure both the quantity and quality of annotations, we could only select four important and representative VL tasks. Through in-depth, fine-grained analysis of the results on these tasks, we have found strong evidence that existing VLMs lack knowledge closely related to Chinese culture and proposed fine-tuning on Chinese cultural VL data as a solution to enhance VLMs' Chinese cultural VLU capabilities. Therefore, we believe that CVLUE is a solid step in the development of Chinese cultural VL benchmarks and hope it will inspire the creation of more extensive and comprehensive Chinese cultural VL datasets.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *CoRR*, abs/2405.17475.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Michael Grubinger, Paul Clough, Henning M "uller, and Thomas Deselaers. 2006. The iapr benchmark: A new evaluation resource for visual information systems. In *Language Resources and Evaluation*, pages 13–23, Genoa, Italy.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *NeurIPS*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 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*, pages 1549–1557. ACM.

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: revealing global cultural perception in language models through natural language prompting. *CoRR*, abs/2404.10199.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 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*, pages 271–275. ACM.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Trans. Multim.*, 21(9):2347–2360.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: A chinese multimodal pretrainer. *CoRR*, abs/2103.00823.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. 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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, Guanhui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In *NeurIPS*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

George A. Miller. 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*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8253–8280. Association for Computational Linguistics.

Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. 2022. Dureader$_{\text{vis}}$: A chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1338–1351. Association for Computational Linguistics.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 171–181. The Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NORMAD: A benchmark for measuring the cultural adaptability of large language models. *CoRR*, abs/2404.12464.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *CoRR*, abs/2406.05967.

Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. 2022. Climb: A continual learning benchmark for vision-and-language tasks. In *NeurIPS*.

Pierre Stock and Moustapha Cissé. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 504–519. Springer.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Bingning Wang, Feiyang Lv, Ting Yao, Jin Ma, Yu Luo, and Haijin Liang. 2022. Chiqa: A large scale image-based real-world question answering dataset for multi-modal understanding. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 1996–2006. ACM.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *CoRR*, abs/2310.12481.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. 2017. AI challenger : A large-scale dataset for going deeper in image understanding. *CoRR*, abs/1711.06475.

Chunyu Xie, Heng Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, Xiangyang Ji, and Yafeng Deng. 2022. Zero and R2D2: A large-scale chinese cross-modal benchmark and A vision-language framework. *CoRR*, abs/2205.03860.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 417–421. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

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

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5731–5746. Association for Computational Linguistics.

Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11762–11771. IEEE.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 17696–17706. ELRA and ICCL.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. 2022. Vlmbench: A compositional benchmark for vision-and-language manipulation. In *NeurIPS*.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. Vlue: A multi-task benchmark for evaluating vision-language models. *CoRR*, abs/2205.15237.

A Appendix

A.1 Categories and Statistics

Categories in MS-COCO
person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, street sign, stop sign, parking meter, bench, *bird*, *cat*, *dog*, *horse*, sheep, *cow*, elephant, bear, zebra, giraffe, hat, backpack, umbrella, shoe, eyeglasses, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, plate, wine glass, cup, fork, knife, *spoon*, *bowl*, banana, *apple*, sandwich, orange, *broccoli*, carrot, hot dog, pizza, donut, cake, *chair*, couch, potted plant, bed, mirror, dining table, window, desk, toilet, door, *TV*, laptop, *mouse*, remote, keyboard, cell phone, microwave, oven, toaster, sink, *refrigerator*, blender, book, clock, vase, *scissors*, teddy bear, hair drier, toothbrush, hairbrush

Table 5: Object categories in MS-COCO.

The 91 object categories in MS-COCO, a popular English VL dataset and often used as the image source for other English and Chinese VL datasets, are listed in Table 5.

The number of annotated objects per category for all 92 categories is shown in Figure 7.

A.2 Data Annotation

In this section, we introduce the detailed data annotation process for all the tasks in CVLUE.

A.2.1 Instance Segmentation

The first stage is the task of segmenting object instances in images of subset A. All the objects belonging to the categories we selected above were manually labelled with bounding boxes.

A.2.2 Image Captioning

The image-text retrieval task includes two subtasks, namely text retrieval (TR), where given an image, the task is to retrieve the corresponding text and image retrieval (IR), where given a text, the task is to retrieve the image. This task aims to evaluate the capability of VLMs to align the semantic space of vision and language representations. The data is annotated via image captioning. Specifically, the annotators were asked to write five different sentences describing each image, which were required to:

- Describe all the important parts of the image.
- Do not describe things that might have happened in the future or past.

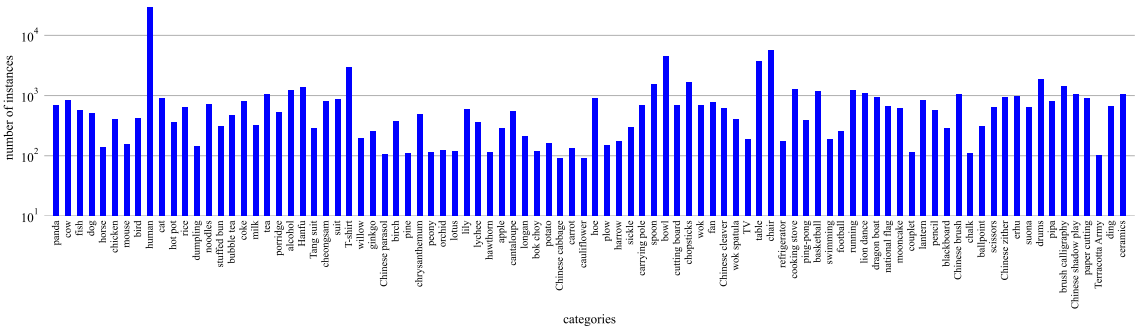


Figure 7: Number of annotated objects per category in CVLUE.

- Do not describe what a person might say.
- Do not name people in the image.
- Contain at least eight characters.
- Contain no more than 30% overlapped characters between each other.

A.2.3 Visual Question Answering

Given an image and a natural language question, the VQA task requires the model to generate or select the corresponding answer in natural language. This task aims to evaluate VLMs' detailed visual understanding and complex reasoning ability. Specifically, the annotators were asked to write three different questions for each image and give the correct answers in short phrases. The questions must: (1) require the image to correctly answer and not be answerable with only commonsense knowledge (e.g., 'What is the book made of?'); and (2) not be too simple that only low-level computer vision knowledge is required to answer them (e.g., 'What colour is the flower?'). The answers must be brief phrases rather than complete sentences. This constraint was added to ensure that the function of the VQA task is distinct from that of the VD task, in which the annotators were required to write complete sentences.

A.2.4 Visual Grounding

Given an image and a natural language referring expression, the VG task requires the model to locate the corresponding object. This task aims to evaluate the VLMs' ability to understand and distinguish objects in images. Specifically, each image was annotated by two annotators, namely A and B. A was asked to write an expression for each object labelled in the instance segmentation stage, distinguishing it from others of the same category.[14] B was then given the expressions one by one and asked to select the corresponding object by clicking on the image. The annotation was regarded as correct only if B correctly selected all the objects.

An important factor that makes this task challenging enough is ensuring that at least two objects of the same category exist in all the images. Otherwise, this task would be degraded into simply distinguishing objects of different categories. Kazemzadeh et al. (2014) built their dataset on images from eixsting ImageCLEF dataset (Grubinger et al., 2006). Therefore, they had no choice but to use images with and without multiple objects of the same category. To deal with this issue, we restrict the number of objects of the same category from the beginning. Specifically, in the collection stage of subset A, we strictly require that only images containing at least two objects of the same category be included. Such categories will be considered as the *main category* of the image. Then, during the VG annotation stage, the annotators were only asked to write expressions for the objects of the images' *main category*. In this way, we guarantee that all the images used in this task contain two or more described objects of the same category, making the task more challenging.

A.2.5 Visual Dialogue

We employ the task of visual dialogue to evaluate the general intelligence of the VLM, ranging from global visual understanding to history memorization and natural language generation. The annotation of the VD task also requires the annotators to work in pairs. One of them was given a caption describing the image from subset B and was required to ask questions about the image to 'imagine the

[14] For images containing more than four objects of the same category, we let the annotator select four objects to annotate.

scene better'. Another annotator was given both the image and the caption and was required to answer the questions based on the image. The conversation will be ended after ten pairs of questions and answers. It was emphasized to the annotators that the questions must be related to concrete objects in the image. Abstract questions concerning reason and meaning were not allowed.

A.3 Data Characteristics

A.3.1 Images and Objects

The numbers of annotated categories per image are shown in Figure 8.

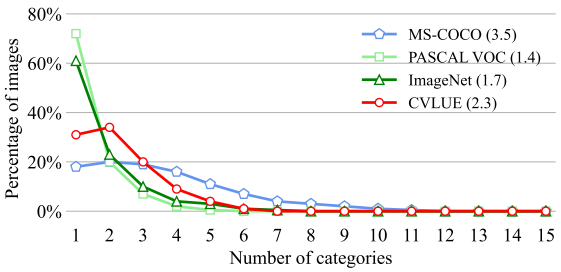


Figure 8: Number of annotated categories per image for CVLUE, MS-COCO, ImageNet Detection and PASCAL VOC (average number of categories are shown in parentheses).

A.3.2 Visual Question Answering and Visual Dialogue

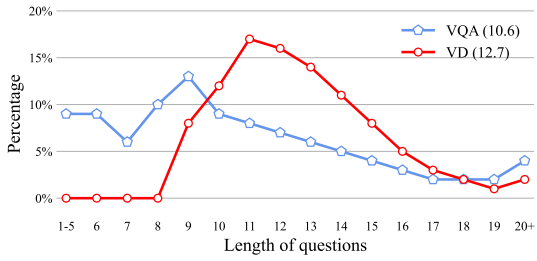


Figure 9: The question length distribution of VQA and VD in CVLUE (average lengths in parentheses).

To illustrate the difference between VQA and VD tasks, we report their distribution of question and answer lengths in Figure 9 and Figure 10, respectively. The question length distribution shows that VD has longer questions than VQA on average. The difference becomes more evident in the answer length distribution, where answers in VQA are all short phrases, while VD has much longer answers.

This difference reflects the distinct motivation of these two tasks. With VQA, we want the model

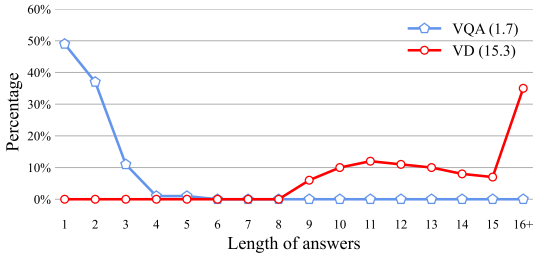


Figure 10: The answer length distribution of VQA and VD in CVLUE (average lengths in parentheses).

to focus more on detailed visual understanding and complex reasoning. With VD, however, we want to evaluate VLMs' general intelligence, including global visual understanding, history memorization, and natural language generation. We also count the number of sentences containing pronouns (e.g., 'he', 'she', 'it', etc.) and find that 43% questions, 32% answers and almost all (93%) dialogues in VD contain at least one pronoun. In contrast, only 1% of sentences in VQA contain pronouns. This means that the VD task also requires the capability to overcome coreference ambiguities, which is not strictly required by VQA.

To the best of our knowledge, there has not been any similar Chinese VD dataset. So, we make a rough comparison between the VD dataset in CVLUE and its English counterpart, the Visdial 1.0 dataset (Das et al., 2017). We focus on the answers and find that the two most frequent answers for Visdial 1.0 are 'no' and 'yes', constituting 21.3% and 19.2% of the total answers, respectively. For our VD dataset, the two most frequent answers are '这是一个女人/男人' (This is a woman/man), constituting only 0.1% and 0.07% of the total answers, respectively. Overall, Visdial 1.0 has 1,232,870 answers of 337,527 different types, while our VD dataset contains 97,550 answers of 93,308. The average answer lengths are 2.9 words for Visdial 1.0 and 15.3 characters for our VD dataset. This comparison shows our VD dataset's superiority regarding the answers' richness and complexity.

A.4 CVLUE Examples

A.4.1 Image-Text Retrieval

The 5 captions for ITR example 1 in Figure 11 are:

- 地面上有两只舞狮 (There are two dancing lions on the ground)
- 右边红色的舞狮横着站在地上，旁边还有一只黄色的舞狮 (On the right, a red dancing

Figure 11: Image for ITR example 1.

lion is standing horizontally on the ground, with a yellow dancing lion beside it)

- 有一只黄色的舞狮和一只红色的舞狮站在超市前边 (A yellow dancing lion and a red dancing lion are standing in front of a supermarket)
- 黄色的舞狮站起来了，旁边有另一只舞狮看着它，周围还有一些看客 (The yellow dancing lion is standing up, with another dancing lion watching it, surrounded by some spectators)
- 中国珠宝的店铺前有一个喜庆的拱门，前面有几只舞狮正在表演节目 (In front of a Chinese jewellery store, there is a festive archway, and several dancing lions are performing in front of it)

Figure 12: Image for ITR example 2.

The 5 captions for ITR example 2 in Figure 12 are:

- 肉摊上的人在往塑料袋里面装肉，桌子上有菜刀 (The person at the meat stall is putting meat into a plastic bag, and there is a Chinese cleaver on the table)
- 电子秤旁边放着几块肉和一些菜刀，有个人在摊位前选肉 (Next to the electronic scale are some pieces of meat and Chinese cleavers, and a person is selecting meat at the stall)
- 一块圆形菜板上放着一些碎肉和一把菜刀，摊主手里提着塑料袋 (On a round cutting board, there are some pieces of meat and a Chinese cleaver, and the vendor is holding a plastic bag)
- 两块长方形桌子拼在一起，桌子上边有菜刀，下方有几个泡沫盒，有一条鱼在摊主的脚下 (Two rectangular tables are joined together, with Chinese cleavers on top and several foam boxes underneath, and there is a fish at the vendor's feet)
- 卖肉的摊位前有人经过，摊主的前面有菜刀和几块肉，选肉的人正用手指着其中一块 (Someone is passing by the meat stall; in front of the vendor are Chinese cleavers and pieces of meat, and the person selecting meat is pointing at one of the pieces)

Figure 13: Image for ITR, VQA and VG example 3.

The 5 captions for ITR example 3 in Figure 13 are:

- 桌上放着两口火锅和一些火锅食材 (There are two hot pots and some hot pot ingredients on the table)
- 两口火锅的下面放着用杯子装着的蔬菜和肉类 (Below the two hot pots, there are cups filled with vegetables and meat)

- 一些蔬菜和肉类的火锅食材被放在火锅的旁边，火锅下面放着加热灶 (Some vegetables and meat for the hot pot are placed beside the hot pot, and heating stoves are placed under the hot pots)
- 一口鸳鸯火锅和一口麻辣火锅放在了两个加热炉上，锅里面还放着一些食材 (A divided hot pot and a spicy hot pot are placed on two heating stoves, with some ingredients inside the pots)
- 装有食物的搪瓷杯子摆放在托盘上，食材旁边有两口火锅，它们分别放在了两台加热炉上 (Enamel cups filled with food are placed on a tray, and there are two hot pots beside the ingredients, each on a separate heating stove)

A.4.2 Visual Question Answering

Figure 14: Image for VQA Example 1.

The 3 question-answer pairs for VQA example 1 in Figure 14 are:

- Q: 戴帽子的人是在下台阶还是上台阶? (Is the person wearing a hat going down the steps or up the steps?) A: 下台阶 (Going down the steps)
- Q: 有几根担杖? (How many carrying poles are there?) A: 2
- Q: 两根担仗中间的人是坐着还是站着? (Is the person between the two carrying poles sitting or standing?) A: 坐着 (Sitting)

Figure 15: Image for VQA Example 2.

The 3 question-answer pairs for VQA example 2 in Figure 15 are:

- Q: 当前是什么季节? (What season is it currently?) A: 冬季 (Winter)
- Q: 右侧大熊猫是什么姿势? (What is the posture of the panda on the right?) A: 坐着 (Sitting)
- Q: 坐着的大熊猫数量与站立的大熊猫数量相减等于几? (What is the difference between the number of sitting pandas and standing pandas?) A: 0

The 3 question-answer pairs for VQA example 3 in Figure 13 are:

- Q: 圆形切片的是什么食材? (What ingredient is the round slice?) A: 藕 (Lotus root)
- Q: 火锅的口味相同吗? (Are the flavors of the hot pots the same?) A: 不相同 (No)
- Q: 鸳鸯锅里褐色食材是什么? (What is the brown ingredient in the divided hot pot?) A: 香菇 (Shiitake mushrooms)

A.4.3 Visual Grounding

The 4 referring expressions for VG example 1 in Figure 16 are:

- 1.扇子上有汉字的剪纸 (The paper cutting on the fan in the shape of Chinese characters)
- 2.扇子上有三朵花形状的剪纸 (The paper cutting on the fan in the shape of three flowers)
- 3.有老虎形状的红色剪纸 (The red paper cutting in the shape of a tiger)

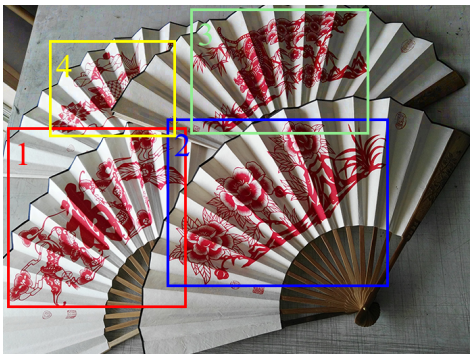


Figure 16: Image for VG Example 1.

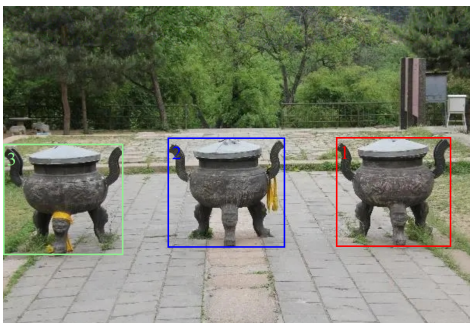


Figure 17: Image for VG Example 2.

- 4.扇子上是荷叶和鱼形状的剪纸 (The paper cutting on the fan in the shape of lotus leaves and fish)

The 3 referring expressions for VG example 2 in Figure 17 are:

- 1: 鼎身没有系黄色绸带的一只鼎 (The ding without a yellow silk ribbon tied around its body)
- 2: 位于中间的一只鼎 (The ding in the middle)
- 3: 鼎脚上系了一根黄色绸带的一只鼎 (The ding with a yellow silk ribbon tied around its leg)

The 2 referring expressions for VG example 3 in Figure 13 are:

- 1: 汤汁上飘有许多干辣椒的那个火锅 (The hot pot with many dried chili peppers floating on the broth)
- 2: 装有香菇的那个鸳鸯火锅 (The divided hot pot with shiitake mushrooms)

Figure 18: Image for VD Example 1.

A.4.4 Visual Dialogue

The caption and 10 rounds of dialogue for VD example 1 in Figure 18 are:

- Caption: 戴着黑色帽子的男人在拉二胡 (A man wearing a black hat is playing the erhu)
- Q: 这个男人穿着什么样的衣服? (What kind of clothes is this man wearing?) A: 这个男人穿着黑色的长袍 (This man is wearing a black robe)
- Q: 这个男人是在室内还是室外? (Is this man indoors or outdoors?) A: 这个男人是在室外 (This man is outdoors)
- Q: 这个男人周围都有什么? (What is around this man?) A: 这个男人旁边有音响和水杯, 背后有门和墙 (Next to this man, there are a speaker and a water cup, and behind him, there is a door and a wall)
- Q: 这个男人背后的墙是什么样式的? (What is the style of the wall behind this man?) A: 这个男人背后的墙是灰色的砖块砌成的 (The wall behind this man is made of grey bricks)
- Q: 这个男人背后的门是什么颜色的? (What colour is the door behind this man?) A: 这个男人背后的门是棕黄色的 (The door behind this man is brownish-yellow)
- Q: 男人背后的门是什么材质的? (What material is the door behind the man made of?) A: 男人背后的门是木头制作的 (The door behind the man is made of wood)
- Q: 男人背后的门是打开的吗? (Is the door behind the man open?) A: 男人背后的门是打开的 (The door behind the man is open)

- Q: 男人旁边的音响是什么颜色的? (What colour is the speaker next to the man?) A: 男人旁边的音响是黑色的 (The speaker next to the man is black)
- Q: 男人旁边的水杯是什么颜色的? (What colour is the water cup next to the man?) A: 男人旁边的水杯是银白色的 (The water cup next to the man is silver)
- Q: 这个水杯是盖着的吗? (Is the water cup covered?) A: 这个水杯是盖着的 (The water cup is covered)

Figure 19: Image for VD Example 2.

The caption and 10 rounds of dialogue for VD example 2 in Figure 19 are:

- Caption: 砖墙前倚靠着一位旗袍女子 (A woman in a cheongsam leaning against a brick wall)
- Q: 图中的砖墙是什么颜色的呢？ (What colour is the brick wall in the picture?) A: 图中的砖墙是浅灰色的砖墙。 (The brick wall in the picture is light grey.)
- Q: 图中的女子穿着一件什么颜色的旗袍呢？ (What colour is the cheongsam the woman is wearing?) A: 图中的女子穿了一件浅绿色的旗袍。 (The woman in the picture is wearing a light green cheongsam.)
- Q: 图中女子穿的旗袍是长袖还是短袖的呢？ (Is the cheongsam the woman is wearing long-sleeved or short-sleeved?) A: 图中的女子穿的旗袍是短袖的。 (The cheongsam the woman is wearing is short-sleeved.)
- Q: 图中女子的动作是什么样子的呢？ (What is the woman doing in the picture?) A: 女子倚靠着砖墙，右手抚摸着耳朵眼睛看着镜头。 (The woman is leaning against the brick wall, touching her ear with her right hand and looking at the camera.)
- Q: 图中女子是在室内还是室外呢？ (Is the woman indoors or outdoors in the picture?) A: 女子是在室外，她的身后还有很多花。 (The woman is outdoors, with many flowers behind her.)
- Q: 女子身后的花是什么颜色的呢？ (What colour are the flowers behind the woman?) A: 女子身后的花有浅粉色的和橘黄色的。 (The flowers behind the woman are light pink and orange.)
- Q: 图中女子看上去年龄有多大呢？ (How old does the woman in the picture look?) A: 图中的女子看上去很年轻，二十来岁。 (The woman in the picture looks very young, in her twenties.)
- Q: 图中除了女子，还有其他的人吗？ (Are there any other people in the picture besides the woman?) A: 图中只有女子一个人，没有其他的人。 (There is only the woman in the picture, no one else.)
- Q: 图中的天气情况怎么样呢？ (What is the weather like in the picture?) A: 图中阳光明媚，是一个晴天。 (The weather in the picture is sunny and bright.)
- Q: 图中的女子身上有什么装饰品吗？ (Is the woman wearing any accessories in the picture?) A: 图中的女子头上带了一个浅棕色的发饰。 (The woman in the picture is wearing a light brown hair accessory.)

The caption and 10 rounds of dialogue for VD example 3 in Figure 20 are:

- Caption: 屋内坐着一个穿着汉服的女生 (A girl in Hanfu sitting indoors)
- Q: 她的旁边有什么东西？ (What is next to her?) A: 有一个桌子和毛笔架。 (There is a table and a Chinese brush rack.)

Figure 20: Image for VD Example 3.

- Q: 毛笔架上面有毛笔吗？ (Are there brushes on the brush rack?) A: 是的，上面挂着五支毛笔。 (Yes, there are five Chinese brushes hanging on it.)
- Q: 她穿的汉服是什么颜色的？ (What colour is the Hanfu she is wearing?) A: 浅绿色的有金色的图案。 (It is light green with golden patterns.)
- Q: 她的手在什么位置？ (Where are her hands?) A: 右手握着毛笔，左手放在纸上面。 (Her right hand is holding a Chinese brush, and her left hand is on the paper.)
- Q: 画纸上面有镇纸吗？ (Is there a paperweight on the drawing paper?) A: 没有，桌子上只有画纸和毛笔架。 (No, there is only drawing paper and the Chinese brush rack on the table.)
- Q: 她的旁边没有其他人吗？ (Is there no one else next to her?) A: 是的，只有她一个人。 (Yes, she is alone.)
- Q: 毛笔架上面的毛笔都长得一样吗？ (Do all the brushes on the brush rack look the same?) A: 不是，每个毛笔的样式都不一样。 (No, each Chinese brush is different.)
- Q: 桌子是什么颜色的？ (What colour is the table?) A: 是深红色的与黑色混杂的桌面。 (It is a dark red and black mixed tabletop.)
- Q: 她的后面有什么东西？ (What is behind her?) A: 是一些木质的架子，原木的颜色。 (There are some wooden shelves, natural wood colour.)
- Q: 她的头上有什么装饰品？ (Is she wearing any accessories on her head?) A: 有绿色的和白色的小花。 (She has small green and white flowers in her hair.)

A.5 Fine-tuning Experimental Setups

In the fine-tuning setting, all tasks use the AdamW optimizer with a weight decay of 0.05 and the cosine learning rate scheduler. We use the default image resolution for each of the baseline models. Other hyper-parameters are listed in Table 6. In the fine-tuning setting, during the inference stage of VQA, we constrain the decoder to only generate from candidates computed in the training and valid set. The models were fine-tuned on 8 V100s.

Task	init LR	batch size	resolution	#epoch
ITR	$3e^{-5}$	128	384×384	10
VQA	$3e^{-5}$	128	768×768	5
VG	$1e^{-5}$	128	384×384	10
VD	$3e^{-5}$	128	384×384	5

Table 6: Hyper-parameters used in the fine-tuning setting. init LR stands for initial learning rate.

A.6 Experimental Results

The data splits of the English VL datasets we used are shown in Table 7.

Task	|Train|	|Valid|	|Test|
COCO (5K)	82,783	5,000	5,000
VQA-v2	82,783	40,504	81,434
RefCOCOg	21,899	1,300	2,600
Visdial 1.0	123,287	2,064	8,000 (QA pairs)

Table 7: Data splits (in terms of image numbers if not explicitly specified) of the English VL datasets we used.

The full experimental results are shown in Table 8.

A.7 Prompts for the Zero-Shot Setting

A.7.1 Visual Question Answering

In the VQA task, we use the prompts '用尽量简洁的数字或中文短语回答以下问题：[question]' for Chinese and 'Answer the question with only an Arabic figure or a phrase: [question]' for English, where [question] denotes the question in VQA.

A.7.2 Visual Grounding

In the VG task, we use the prompts '框出图中[expression]的位置' for Chinese and '<ref>[expression]</ref><box>' for English, where [expression] denotes the referring expression in

Tasks	Dataset	Metrics	Fine-tuning		Zero-shot		
			CCLM 522M	X^2VLM 422M	QwenVL 7B	QwenVL-Chat 7B	mPLUG-Owl2 7B
TR	COCO (5K)	R@1	77.7	80.1	-	-	-
		R@5	94.2	95.3	-	-	-
		R@10	97.1	97.6	-	-	-
	CVLUE	R@1	49.9	54.8	-	-	-
		R@5	75.2	79.5	-	-	-
		R@10	82.8	86.8	-	-	-
IR	COCO (5K)	R@1	60.5	63.8	-	-	-
		R@5	84.3	86.1	-	-	-
		R@10	90.7	91.8	-	-	-
	CVLUE	R@1	32.0	36.6	-	-	-
		R@5	58.3	63.4	-	-	-
		R@10	69.6	73.6	-	-	-
VQA	VQA-v2 (test-std)	Acc	63.7	75.5	78.0	67.9	79.2
	CVLUE	Acc	58.5	53.0	29.9	39.8	20.4
VG	RefCOCOg	IoU	70.4	79.9	78.0	80.1	-
	CVLUE	IoU	39.1	48.8	36.8	40.4	-
VD	Visdial 1.0	R@1	42.4	41.5	36.0	37.5	37.2
		R@5	64.4	59.7	50.0	51.8	52.4
		R@10	72.5	67.7	55.6	57.6	59.4
	CVLUE	R@1	32.2	27.6	24.8	26.5	25.8
		R@5	46.6	41.0	34.9	35.9	38.3
		R@10	53.3	47.8	39.9	40.2	45.3

Table 8: Results of baseline VLMs. R@1, R@5 and R@10 denote the recall in the top 1, 5 and 10 predictions, respectively. Acc denotes the accuracy, and IoU stands for the average intersection over union. For each compared model, we also report the number of parameters.

VG, <ref>, </ref> and <box> are special tokens in the Qwen-VL model.

A.7.3 Visual Dialogue

In the VD task, we use the prompts '描述: [caption] 对话历史: [history] 根据图片描述和对话历史用一句话回答以下问题. 问题: [question] 答案:' for Chinese and 'Context: [caption] History: [history] Answer the question with one sentence based on the context and dialogue history. Question: [question] Answer:' for English. [caption] denotes the caption describing the image in VD, [history] denotes the dialogue history, which is also in the format of question-answer pairs, and [question] denotes the current question to be answered in this round of dialogue.

Since the VD task is to rank the 100 answer candidates given the dialogue history and current question, we could not directly apply the generative VLMs in such a situation. Therefore, we concatenate each answer candidate with the dialogue history and the current question and use the VLM to calculate their probabilities, eventually ranking all candidate answers based on these probabilities.

A.8 Results on Translated English Test Sets

To ensure translation quality, we used the gpt-4-1106-preview model. The translation examples listed in Table 9 demonstrate that this model can accurately translate texts containing categories closely related to Chinese culture.

Chinese	Translated English
穿蓝色上衣的男人拿着的唢呐	The suona held by the man in the blue shirt
距离岸堤最近的一艘龙舟	The dragon boat closest to the shore
穿着深色唐装的人在拉的二胡	The erhu being played by the person in the dark Tang suit
头扭向一侧且戴着眼镜的女生左手拿的琵琶	The pipa held in the left hand of the female with her head turned to one side and wearing glasses
里面放有许多绿色青菜的那个火锅	The hotpot containing many green vegetables

Table 9: Examples of original Chinese and translated English in the CVLUE VG test set.

Figure 21 shows the results of QwenVL, QwenVL-Chat, and mPLUG-Owl2 on the original Chinese test set and the translated English test set for the CVLUE VQA task. It is worth noting that the mPLUG-Owl2 model exhibits a significant performance gap between the original Chinese and translated English test sets. Analysis of the prediction results reveals that this discrepancy is due to the model's misunderstanding of the prompt. Despite the input prompt explicitly instructing the

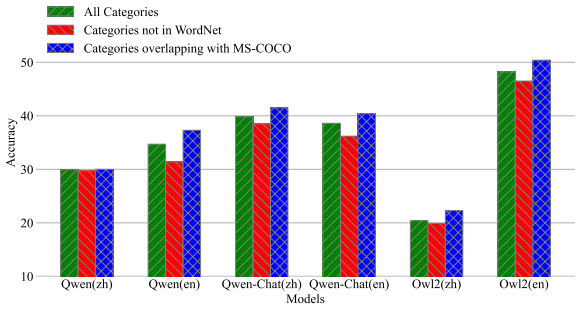


Figure 21: Category group results of QwenVL, QwenVL-Chat and mPLUG-Owl2 on the original Chinese (zh) and translated English (en) CVLUE VQA test set.

model to answer in Chinese (see Appendix A.7.1), 56% of the model's predictions were still in English. Therefore, the model's performance on the Chinese test set does not fully reflect its knowledge of Chinese culture.

A.9 Zero-Shot vs. Fine-Tuninig

The category group results of zero-shot and fine-tuned models on VQA are shown in Figure 22.

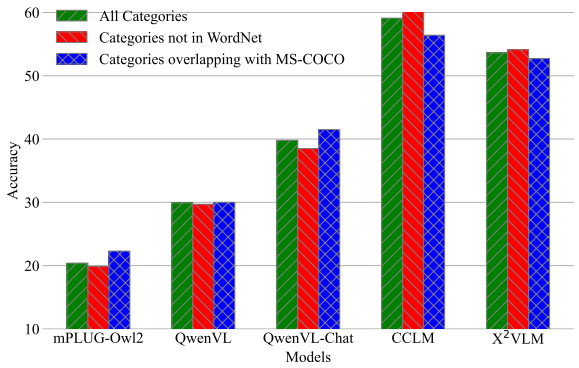


Figure 22: Category group results on CVLUE VQA task.

A.10 Results by Category

The results by category on IR, TR, VQA, VG and VD tasks are shown in Figure 23, 24, 25, 26 and 27, respectively.

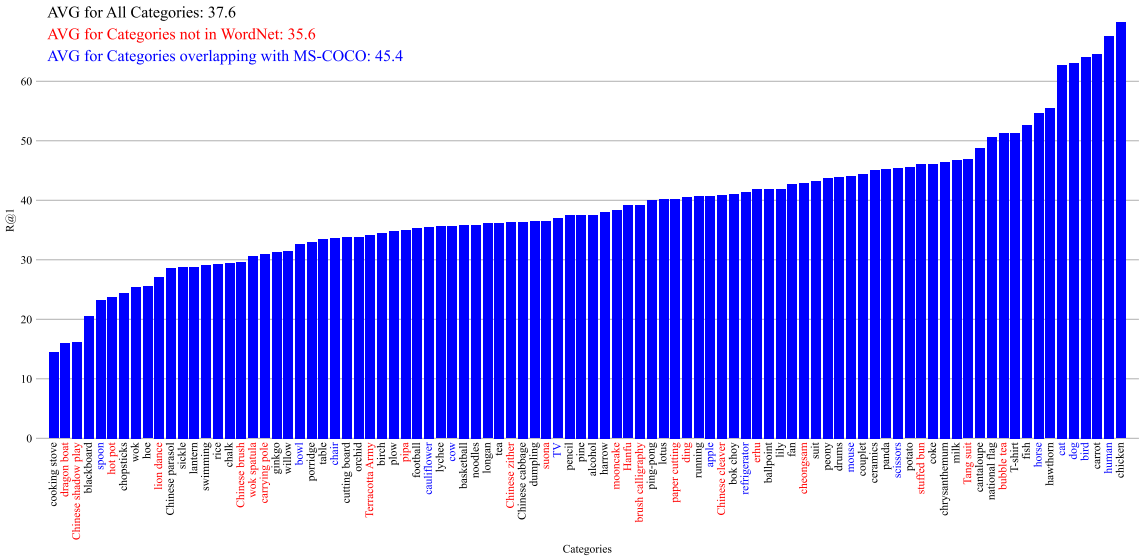


Figure 23: Results of X^2VLM model on the CVLUE IR task, displayed by image category.

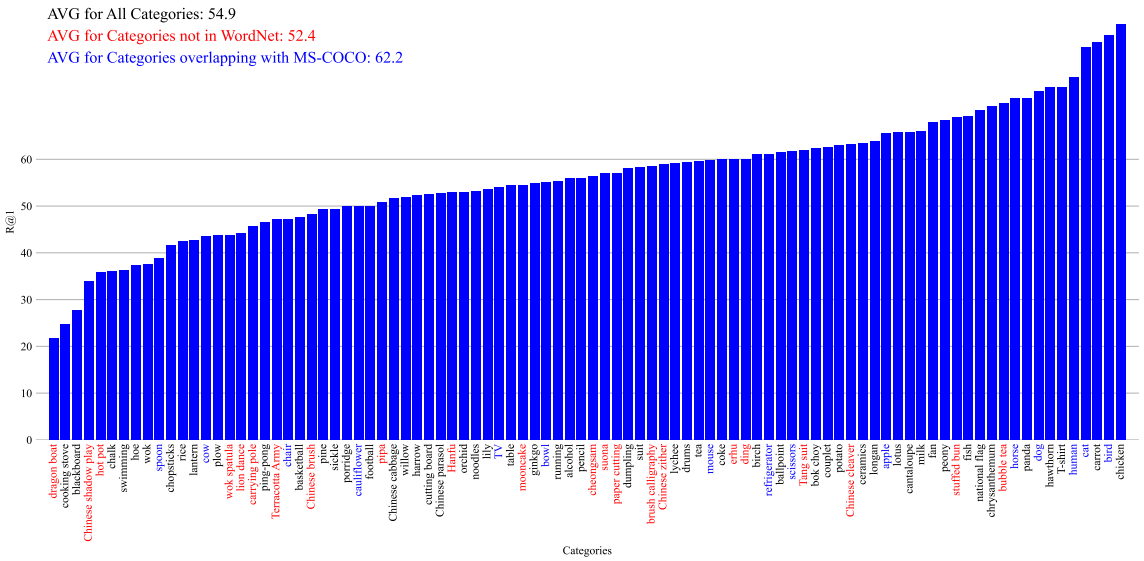


Figure 24: Results of X^2VLM model on the CVLUE TR task, displayed by image category.

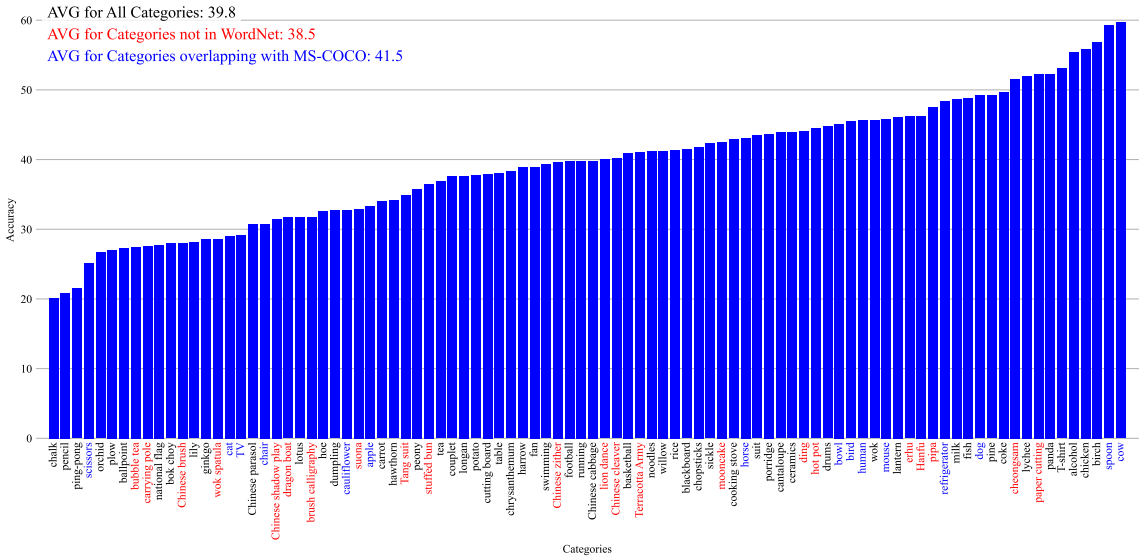


Figure 25: Results of QwenVL-Chat model on the CVLUE VQA task, displayed by image category.

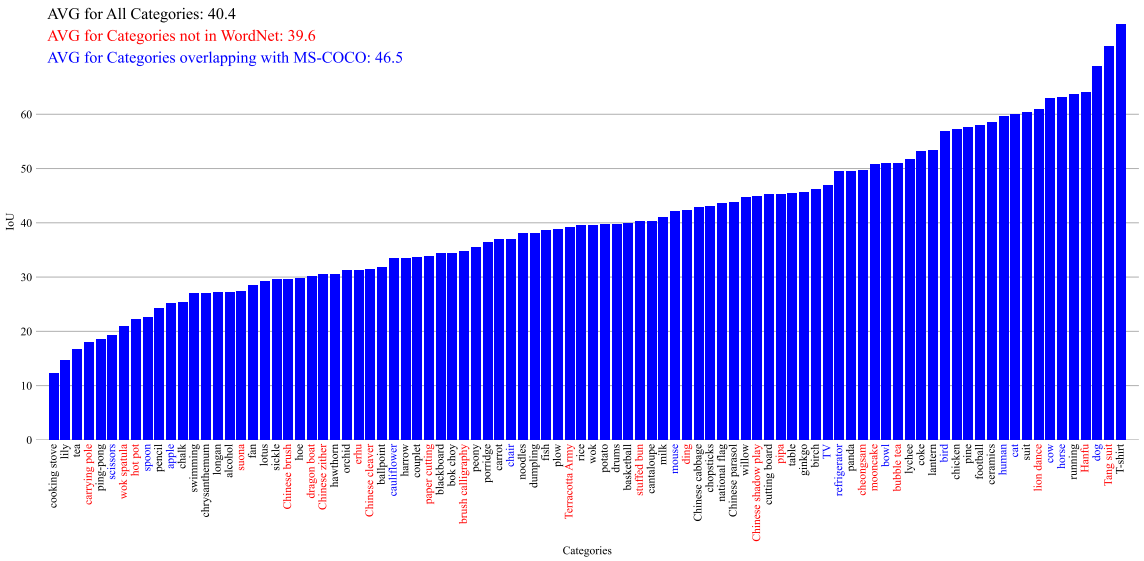


Figure 26: Results of QwenVL-Chat model on the CVLUE VG task, displayed by image category.

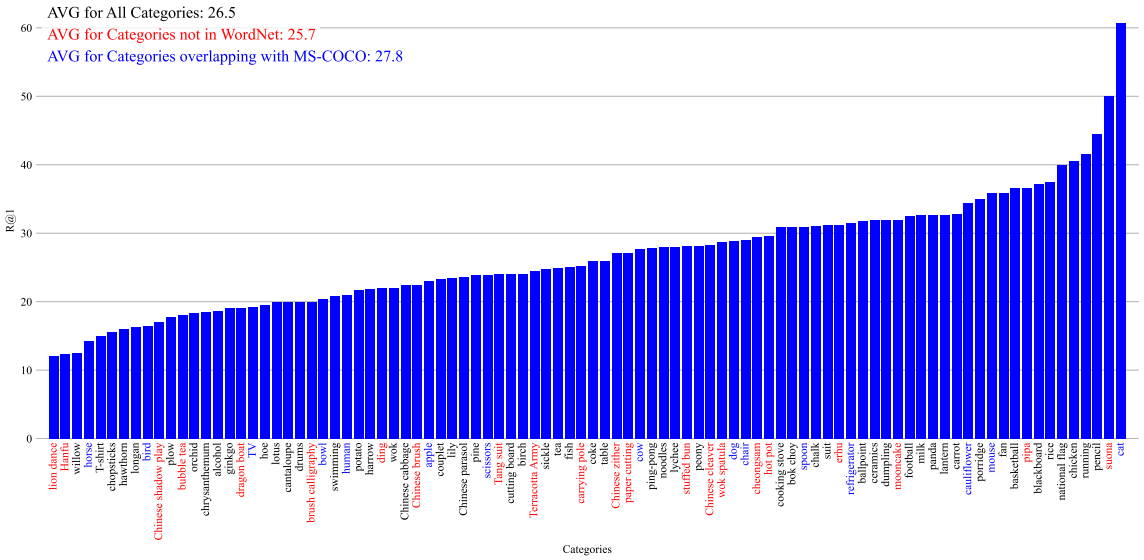


Figure 27: Results of QwenVL-Chat model on the CVLUE VD task, displayed by image category.