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

Anonymous ACL submission

Abstract

The abundance of vision-language (VL) understanding benchmark datasets for English, such as MS-COCO and Flickr30K, has largely facilitated the evaluation of new vision-language models (VLMs) across diverse tasks. However, despite the rapid development of Chinese VLMs, most existing Chinese VL datasets are constructed by re-annotating the images from English VL datasets, limiting the source of images to English-speaking cultures only. Some others are limited to a few fundamental tasks, like image-text retrieval. Such cultural bias and limitation of task types make these datasets unsuitable and inadequate 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, which evaluates a model's VL capability from multiple aspects. 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. [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 (Fellbaum, 2010), 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.

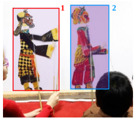
A line of research studies (Shankar et al., 2017; DeVries et al., 2019; Liu et al., 2021) have been conducted to remedy this issue, where one of the most effective ways is to collect and annotate images in other languages and cultures directly. For example, Liu et al. (2021) constructed a multilingual dataset for Multicultural Reasoning over Vision and Language (MaRVL), which contains images collected and annotated by native speakers of 5 typologically diverse languages ranging from Chinese to Turkish. However, only a limited number of images were annotated for each language, and only one VL task was involved in this dataset.

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.

(a) Image Captioning

(b) Visual Question Answering



(c) Visual Grounding

(d) Visual Dialogue

Figure 1: Examples of the images and their annotation for the four tasks in CVLUE. The annotation of image captioning is used for the ITR task.

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. 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[2], 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, where the main objects' categories in the examples are hot pot, dragon boat, shadow puppet and stuffed bun, respectively (all representative in Chinese culture). Among the 92 categories in CVLUE and 91 categories in MS-COCO (commonly used as the image source for English VL datasets), only 15 are overlapped.[3] And the non-overlapped ones are mostly

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

[3]Please refer to Appendix A.1 for the full list of categories in CVLUE and MS-COCO.

representative of Chinese culture. To the best of our knowledge, CVLUE is the most comprehensive Chinese VL benchmark dataset so far.

We believe this dataset can provide a fair and convenient platform for the evaluation of VLMs in Chinese culture and facilitate the evaluation and development of the Chinese VLP. We present a detailed statistical analysis to show the distinct properties and goals of the four tasks involved and benchmark several popular open-source multilingual VLMs on CVLUE and some established English VL datasets to evaluate their VL understanding capability in Chinese and English.

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). A line of work has been conducted to solve this problem. For instance, Yang et al. (2020) proposed to intervene in the data by filtering and re-balancing a subset of categories. Liu et al. (2021) introduced a natural language visual reasoning dataset covering five languages, where the image collection and annotation were driven by native speakers. Unfortunately, only a limited number of data for a single task was provided for each language in their dataset.

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, the recent English VL benchmarks were constructed mainly by exploiting existing VL datasets. However, given the situation of existing Chinese VL datasets, it is undoubtedly much more challenging to build a VL benchmark dataset specifically designed for Chinese. 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. Besides, like all the Chinese VL datasets discussed above, no explicit rule was mentioned in the image collection stage to ensure that the selected images were representative of Chinese culture. Hence, there is a considerable chance that images in such datasets may fail to reflect the actual distribution in Chinese culture. To remedy this issue, we present CVLUE, where the collection of images was entirely driven by Chinese native speakers with explicit constraints to ensure that the source images are representative of Chinese culture. Our benchmark covers four distinct VL tasks, which help to evaluate the general capability of Chinese VLMs from multiple aspects.

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

Task	|Train|	|Valid|	|Test|	Metrics
ITR	25,761	4,248	11,655	R@k
VQA	20,697	3,405	6,046	Acc
VG	15,571	2,548	5,885	IoU
VD	5,748	914	3,093	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.

metrics are summarized in Table 2. In this section, we describe the procedure we devised for image collection and dataset annotation.

3.1 Selection of Object Categories

We first introduce how the object categories are selected. The categories must form a representative set of all categories in Chinese daily life, reflecting the unique characteristics of Chinese culture. The selection of object categories for our dataset is inspired by the Chinese part of MaRVL (Liu et al., 2021), where five native speakers are asked to provide 5-10 specific concepts of 18 semantic fields. The concepts must be 'commonly seen or representative in the speaking population of the language' and 'be physical and concrete'. However, since CVLUE is developed specifically for Chinese, while MaRVL is created for multiple languages, its categories are unsuitable for direct application here.

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 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 select 92 object categories from 15 semantic fields listed in Appendix A.1.

3.2 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 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.[4] Subset B consists of images *containing 3-5 objects of different object categories* and is used for the VD task.[5] 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.

3.3 Quality Control

Given the complexity of the tasks in our dataset, the selection and training of annotators are of great importance and consist of two steps. In the first step, the annotation guidelines for all the tasks were given to the candidates, who were asked to annotate all the tasks on 5 randomly sampled images to evaluate their general capability. Qualified candidates were then categorized into groups for specific tasks based on their performance in the general test. In the second step, annotators in each group were asked to annotate their specific task on 50 randomly sampled images. They were instructed one-on-one by senior annotators well aware of the guidelines until they fully understood them and their annotation was 100% correct.

Annotators who had gone through the above steps were allowed to start annotating. The tasks were batched into packages, and an annotator could not apply for the next package until the current one was finished. The annotation of each package was *first checked by the annotator himself or herself, then checked by a senior inspector, and eventually inspected by four co-authors well aware of the annotation guidelines*. Each package was sampled by 10%-25% for the final inspection by the co-authors, and only those with accuracy higher than 97% could pass it. Otherwise, the package will be returned to the annotator and should be double-checked and corrected. Overall, 41, 108, 44, 26 annotators and 10, 12, 8 and 13 senior inspectors were involved in the annotation procedure of IC, VQA, VG and VD tasks, respectively. The project took six months to complete, with an expenditure of approximately RMB 550,000.

[4]This constraint ensures VG is challenging enough.

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

3.4 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.

3.5 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. Our guidelines for image caption annotation were mainly inspired by Chen et al. (2015). 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.
- 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.

3.6 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. We devised our annotation guidelines for VQA following Antol et al. (2015). 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.

3.7 Visual Grounding

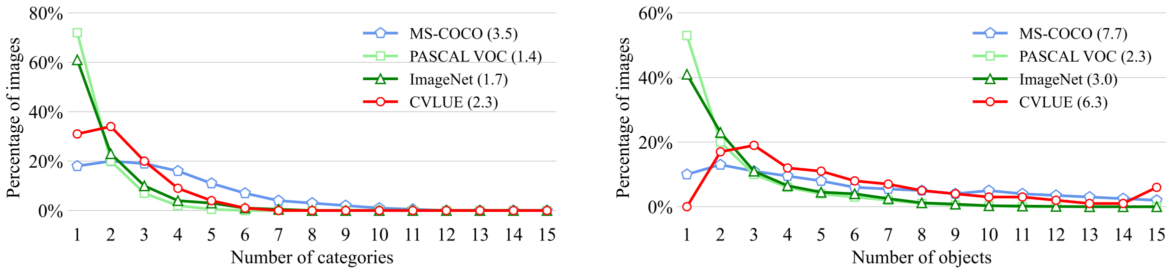
Given an image and a natural language referring expression, the VG task requires the model to locate the corresponding object. This task aims to evaluate the VLMs' ability to understand and distinguish objects in images. The annotation of the VG task is accomplished in a process similar to the Referring Expression Game (Kazemzadeh et al., 2014). 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.[6] 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.

3.8 Visual Dialogue

Inspired by Das et al. (2017), 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

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



(a) Categories per image

(b) Objects per image

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

image to 'imagine the 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.

4 Dataset Analysis

In this section, we extensively analyse all the tasks 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 CVLUE with several popular datasets, including MS-COCO (Lin et al., 2014), ImageNet[7] (Deng et al., 2009) and PASCAL VOC (Everingham et al., 2010). These datasets were designed for various goals. Specifically, MS-COCO was created to detect and segment objects occurring in their neural context. ImageNet was focused on capturing a large number of object categories. Eventually, the primary application of PASCAL VOC was to detect objects in natural images. CVLUE, however, is specifically designed to evaluate VLMs comprehensively in Chinese VLU. The numbers of annotated categories and objects per image are shown in Figure 2a and Figure 2b, respectively, which could reflect the amount of contextual information in the images. Our dataset contains 2.3 categories and 6.3 objects annotated per image on average. In contrast, ImageNet and PASCAL VOC only have less than 2 categories and 3 objects per image on average. Another observation is that none of the images in CVLUE contains one object. This is due to the constraint that all the images in subset A should include at least two objects of the same category in the image collection stage.

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

4.2 Image Captions

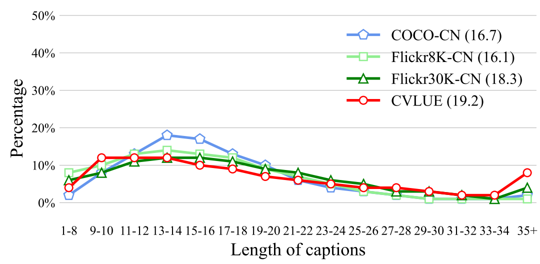


Figure 3: The caption length distribution of CVLUE, COCO-CN, Flickr8K-CN and Flickr30K-CN (average caption lengths are shown in parentheses).

For the image captions used in 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.

4.4 Visual Question Answering and Visual Dialogue

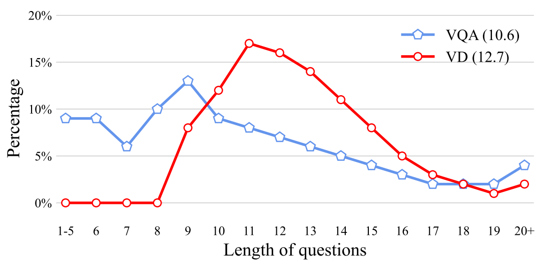


Figure 4: The question length distribution of VQA and VD in CVLUE (average lengths in parentheses).

To illustrate the difference between VQA and VD tasks, we report their distribution of question and answer lengths in Figure 4 and Figure 5, 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.

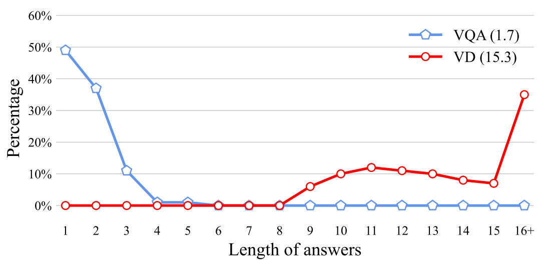


Figure 5: The answer length distribution of VQA and VD in CVLUE (average lengths in parentheses).

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

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).[8]

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

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

Tasks	Dataset	Fine-tuning		Zero-shot		
		CCLM 522M	X^2VLM 422M	QwenVL 7B	QwenVL-Chat 7B	mPLUG-Owl2 7B
TR	COCO (5K)	77.7	80.1	-	-	-
	CVLUE	20.3	23.9	-	-	-
IR	COCO (5K)	60.5	63.8	-	-	-
	CVLUE	15.5	18.0	-	-	-
VQA	VQA-v2 (test-std)	63.7	75.5	78.0	67.9	79.2
	CVLUE	40.6	34.2	25.5	28.1	23.3
VG	RefCOCOg	70.4	79.9	78.0	80.1	-
	CVLUE	36.6	44.3	37.0	39.5	-
VD	Visdial 1.0	42.4	41.5	36.0	37.5	37.2
	CVLUE	31.9	5.4	31.3	33.0	25.6

Table 3: 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.

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.2 and A.3 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 3.[9] All models under the zero-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. 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. It also validates the usefulness of CVLUE in the evaluation of VLMs in Chinese culture.

[9]See Appendix A.4 for full results containing R@5 and R@10 for the TR, IR and VD tasks and detailed discussion.

6 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. 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.

Furthermore, we find that in CVLUE, 17,893 images from subset A are annotated on all three tasks of ITR, VQA and VG, and 9,667 images from subset B are annotated on both the ITR and VD tasks. This could be a beneficial property for future research in joint learning of multiple Chinese VL tasks, which we leave for future study.

7 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.

8 Limitations

To begin with, due to the lack of computational resources, we were unable to test all VLMs on the proposed dataset. Hence, we selected some popular and representative models and conducted experiments under both fine-tuning and zero-shot settings. Also, we couldn't afford to tune hyper-parameters for each model, so we used the same default ones for them all. Therefore, the results reported may not reflect the models' full potential. However, we believe that the current experimental setting is enough to reveal the large performance gap of these strong and popular VLMs between English and Chinese VL datasets. Such observation further validates the usefulness of CVLUE in the comprehensive evaluation of VLMs in Chinese VLU.

Secondly, as mentioned in section 6, a large number of images in CVLUE have been annotated under multiple VL tasks. As this property is beyond the scope of this paper, it is not discussed in detail. However, we believe it will become a valuable and beneficial property in future research, especially in joint learning of multiple Chinese VL tasks.

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.

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.

Christiane Fellbaum. 2010. Harmonizing wordnet and framenet. In *Advances in Natural Language Processing, 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010*, volume 6233 of *Lecture Notes in Computer Science*, page 2. Springer.

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.

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.

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.

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$_{\mathrm{vis}}$: 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.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *CoRR*, abs/1711.08536.

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.

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.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 547–558. ACM.

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.

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

We used 92 object categories from 15 semantic fields in CVLUE, which are shown in Table 4. 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 15 overlapped categories are shown in bold font, where about half of them belong to animals. By comparing the categories of the two datasets, it is easy to find that most of the non-overlapped categories in CVLUE are representative of Chinese culture.

Semantic Fields	**Categories**
Animal	panda, **cow**, fish, **dog**, **horse**, chicken, **mouse**, **bird**, **human**, **cat**
Food	hot-pot, rice, dumpling, noodles, stuffed bun
Beverages	milk-tea, coke, milk, tea, porridge, alcohol
Clothing	Hanfu, Tang suit, chi-pao, 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	hoe, plow, harrow, sickle, carrying pole
Tool	**spoon**, **bowl**, cutting-board, chopsticks, wok, fan, Chinese cleaver, spatula
Furniture	**TV**, table, **chair**, **refrigerator**, cooking stove
Sport	ping-pong, basketball, swimming, football, running
Celebrations	lion-dance, dragon boat, national flag, moon cake, couplet, lantern
Education	pencil, blackboard, brush pen, chalk, ball pen, **scissors**
Instruments	Chinese zither, urheen, suona horn, drums, pipa
Arts	calligraphy, shadow play, paper-cutting, Terracotta Army, tripod, ceramic

Table 4: Object categories in CVLUE.

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 number of annotated objects per category for all 92 categories is shown in Figure 6.

A.2 Prompts for the Zero-Shot Setting

A.2.1 Visual Question Answering

In the VQA task, we use the prompts '只用一个阿拉伯数字或一个词或一个短语回答以下问题：[question]' for Chinese and 'Answer the question with only an Arabic figure or a word or a phrase: [question]' for English, where [question] denotes the question in VQA.

A.2.2 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 VG, <ref>, </ref> and <box> are special tokens in the Qwen-VL model.

A.2.3 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

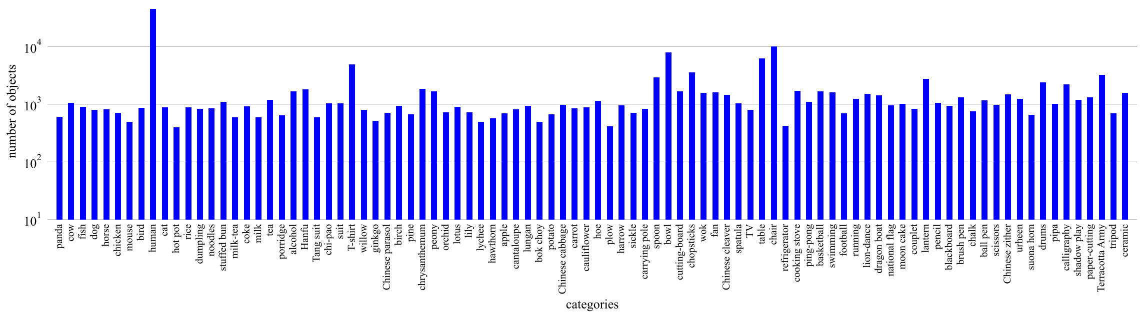


Figure 6: Number of annotated objects per category in CVLUE.

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.3 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.4 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.

We also evaluate the X^2VLM and CCLM models on Flickr8K-CN, a Chinese ITR dataset constructed by re-annotating the Western-culture-biased images from the English Flickr8K dataset. The training, valid, and test sets of Flickr8K-CN contain 6,000, 1,000 and 1,000 images, respectively. The results of X^2VLM and CCLM are shown in Table 8 and 9, respectively. The results show that both models' performance on Flickr8K-CN is even higher than that on COCO (5K). On the contrary, their performance on CVLUE is much lower. We suspect this is because both models were trained on a large number of images with Western cultural biases, which has a similar distribution as the images used in Flickr8K-CN. Meanwhile, images in CVLUE

Tasks	Metrics	Datasets		
		COCO (5K)	Flickr8K-CN	CVLUE
TR	R@1	80.1	92.7	23.9
	R@5	95.3	99.6	46.4
	R@10	97.6	99.7	56.8
IR	R@1	63.8	79.6	18.0
	R@5	86.1	95.5	39.5
	R@10	91.8	98.2	50.6

Table 8: Results of X^2VLM on COCO (5K), Flickr8K-CN and CVLUE.

Tasks	Metrics	Datasets		
		COCO (5K)	Flickr8K-CN	CVLUE
TR	R@1	77.7	89.1	20.3
	R@5	94.2	99.0	41.2
	R@10	97.1	99.8	50.8
IR	R@1	60.5	74.5	15.5
	R@5	84.3	93.6	35.3
	R@10	90.7	97.1	46.0

Table 9: Results of CCLM on COCO (5K), Flickr8K-CN and CVLUE.

Tasks	Dataset	Metrics	Fine-tuning		Zero-shot		
			CCLM 522M	X^2VLM 422M	QwenVL 7B	QwenVL-Chat 7B	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	20.3	23.9	-	-	-
		R@5	41.2	46.4	-	-	-
		R@10	50.8	56.8	-	-	-
IR	COCO (5K)	R@1	60.5	63.8	-	-	-
		R@5	84.3	86.1	-	-	-
		R@10	90.7	91.8	-	-	-
	CVLUE	R@1	15.5	18.0	-	-	-
		R@5	35.3	39.5	-	-	-
		R@10	46.0	50.6	-	-	-
VQA	VQA-v2 (test-std)	Acc	63.7	75.5	78.0	67.9	79.2
	CVLUE	Acc	40.6	34.2	25.5	28.1	23.3
VG	RefCOCOg	IoU	70.4	79.9	78.0	80.1	-
	CVLUE	IoU	36.6	44.3	37.0	39.5	-
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	31.9	5.4	31.3	33.0	25.6
		R@5	46.1	15.3	43.7	45.3	37.1
		R@10	52.6	22.7	49.8	50.9	43.7

Table 10: 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.

are collected under strict constraints, ensuring that they are representative of Chinese culture. This also suggests that existing Chinese VL datasets constructed on top of Western-culture-biased images from English VL datasets are not adequate enough for the evaluation of VLMs' actual VLU capability in Chinese culture.

The full experimental results are shown in Table 10. The performance of X^2VLM on the CVLUE VD task is extremely low. There might be two possible reasons for this. First, as discussed in section 4.4, the answers to be predicted in the CVLUE VD task are more complicated than those in Visdial 1.0, and the low model capacity of X^2VLM might have limited its performance on the CVLUE VD task. Therefore, its performance is much lower than the three large-scale VLMs from the zero-shot setting. Secondly, as shown in Table 2 and Table 7, the VD task in CVLUE has much less training data than Visdial 1.0. Therefore, X^2VLM could neither obtain enough information through fine-tuning in the CVLUE VD task as it did in the Visdial 1.0 dataset.