M2BIT: A Multi-Modal Bilingual Instruction Tuning Dataset for Visual Language Models

Anonymous ACL submission

Abstract

In this paper, we introduce the Multi-Modal Bilingual Instruction Tuning dataset (M2BIT), specifically designed to enhance the performance of vision language models (VLMs). Our M2BIT dataset is one of the largest multi-modal instruction tuning datasets available, covering 40 diverse vision-language tasks in both English and Chinese. It comprises 2 million instances, each accompanied by 400 manually written task instructions. With a carefully curated annotation process, we strive to elevate the quality of response, thereby enriching the user experience while minimizing the generation of potential hallucinations. To validate the efficacy of M2BIT, we train a VLM known as Ying-VLM using this dataset, delving into the impact of instruction tuning across diverse languages and modalities. Upon comparing it with strong VLM baselines, Ying-VLM demonstrates superior performance on complex knowledge vision question answering tasks. Moreover, it exhibits a lower propensity for hallucination, displays greater generalization capabilities to previously unseen video tasks, and better comprehends novel instructions in Chinese. We will open-source the M2BIT dataset and trained models to facilitate future research.

1 Introduction

Following the substantial success of ChatGPT (OpenAI, 2022), the interest in designing a versatile intelligent assistant that can understand and interact with the multi-modal world has surged. The potential of transforming Large Language Models (LLMs) into powerful Vision Language Models (VLMs) has been demonstrated by further training on image-text pairs or implementing specialized visual instruction tuning (Zhu et al., 2023; Liu et al., 2023). This enhancement allows LLMs to *see* the world, offering promising capabilities that could significantly assist individuals with disabilities (OpenAI, 2023).

Dataset	Tasks	LAN	Samples	Ins. / Task
MiniGPT4	N / A	En	5K	N / A
LLaVA	3	En	158K	N / A
MultiModalGPT	3	En	6K	5
MultiInstruct	26	En	235K	5
InstructBLIP	28	En	1.6M	9.7
M2BIT (Ours)	40	En, Zh	2M	10

Table 1: Comparison of different multi-modal instruction tuning datasets. Ins. denotes Task Instruction and N / A means instructions are artificially generated. Our M2BIT is one of the largest datasets, covering 40 tasks in English and Chinese.

The quality of the visual instruction tuning dataset is pivotal in the development of VLMs, as indicated by findings in LLMs (Zhou et al., 2023). Recent research efforts in this area can be broadly grouped into two categories. The first stream of research strives to augment existing academic vision-language (V+L) datasets with manually written task instructions (Xu et al., 2022; Dai et al., 2023). Although VLMs trained on these datasets exhibit notable performance on academic benchmarks, they often generate responses that are excessively terse and blunt (Chen et al., 2023a). This brevity, a typical characteristic of academic datasets, compromises the user experience during interactions. The second stream of research employs image annotation tools to generate textual descriptions of original images, subsequently leveraging models like ChatGPT/GPT-4 to create dialog-style instruction tuning datasets (Zhu et al., 2023; Liu et al., 2023; Zhao et al., 2023a).[1] However, a pitfall of training VLMs with these pseudo-grounded dialogs is the risk of exacerbating LLMs' hallucination problem, possibly resulting in inconsistent image descriptions featuring non-existent objects. Besides, current studies mainly focus on English tasks, which limits the investigation of the cross-lingual effects of instruction tuning.

[1] GPT-4 Vision API is unavailable at the submission time.

In this paper, we introduce M2BIT, a **M**ulti-**M**odal **B**ilingual **I**nstruction **T**uning dataset, which leverages valuable academic benchmarks and the capabilities of ChatGPT for dialog style enhancement. M2BIT is meticulously constructed in three stages: (1) Task Selection and Manual Instruction Writing: Our dataset consists of diverse tasks, including traditional image-text tasks like visual question answering and image captioning, as well as video-related tasks such as video question answering. Annotators are instructed to review the dataset paper thoroughly and craft 10 unique instructions for each task. We further incorporate Chinese vision-language datasets with corresponding Chinese instructions, resulting in a comprehensive compilation of 40 diverse tasks and 400 instructions. (2) Data Format Unification: We ensure that all tasks within our dataset adhere to a unified vision-to-text format. This format comprises four fields: *images*, *instruction*, *inputs* and *outputs*. Additional information, such as bounding box details, is embedded within the images, and short answers are rephrased using ChatGPT while incorporating contextual information, where available. (3) Quality Check: For quality control, we assign an extra annotator to each task to review 20 examples from each split of every dataset. A task is considered complete only after the annotator has verified the accuracy and consistency of the images, instructions, inputs, and outputs for each instance. As demonstrated in Table 1, M2BIT is one of the largest multi-modal instruction tuning datasets regarding the number of instructions and samples, covering diverse tasks in English and Chinese.

To substantiate the effectiveness of the M2BIT dataset, we develop Ying-VLM, merging the capabilities of a vision encoder, BLIP-2 (Li et al., 2023a), with Ziya-13B (Zhang et al., 2022), which is a derivative of LLaMA (Touvron et al., 2023). We leverage the proven methodology of incorporating visual tokens as prefix prompts in LLMs and utilize a two-stage training regime. The initial stage aligns vision features with text embeddings via image captioning on LAION-400M (Schuhmann et al., 2021), while the second stage enhances the model by performing instruction tuning on M2BIT.

Our evaluation of the instruction tuning effect with M2BIT is threefold. Firstly, we evaluate Ying-VLM on knowledgeable VQA (KVQA) tasks, including OK-VQA, A-OKVQA, and a held-out dataset, ViQuAE. These tasks, which demand VLM's comprehension of image context and reasoning with the knowledge acquired by LLMs, have gained wide recognition as benchmarks for evaluating VLMs (Dai et al., 2023; Bai et al., 2023). Secondly, we conduct a hallucination evaluation on image captioning, following the methodology by Li et al. (2023b). Lastly, we perform zero-shot transfer evaluations on Chinese V+L tasks and video-language tasks to scrutinize the cross-language/modal effect of instruction tuning. The experimental results highlight that Ying-VLM surpasses potent baseline VLMs in KVQA tasks, is less susceptible to hallucination than models trained with generated pseudo-grounded dialogs, and demonstrates enhanced generalization capabilities when confronted with unseen video and cross-lingual tasks. These results underscore the potential of our proposed M2BIT dataset in constructing robust VLMs and investigating instruction tuning effects across languages and modalities.

2 M2BIT: A Multi-Modal Bilingual Instruction Tuning Dataset

In this section, we introduce the M2BIT dataset by first elaborating the task coverage (§ 2.1), followed by the annotation process details (§ 2.2). In § 2.3, we present the dataset format and the statistics of the crafted dataset instructions.

2.1 Task Coverage

Our dataset compiles diverse tasks of vision-language tasks, including:

Captioning This task aims to produce descriptions of the given images according to different needs. We include MS COCO (Lin et al., 2014) (the Karpathy split) for generic image descriptions. TextCaps (Sidorov et al., 2020) requires models to capture the text presented in the image and generate captions accordingly. Image-Paragraph-Captioning (Krause et al., 2017) focuses on generating detailed descriptions for images.

Reasoning This task evaluates specific reasoning capabilities. We incorporate CLEVR (Johnson et al., 2017) and NLVR (Suhr et al., 2017) for spatial reasoning, Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) for commonsense reasoning, Visual MRC (Tanaka et al., 2021) for reading comprehensive over images, and Winoground (Thrush et al., 2022) for fine-grained semantics reasoning over text descriptions and image contents.

Visual Question Answering (VQA) This is the most widely studied multi-modal task, which requires the model to answer a given question based on the image correctly. Tasks include VQA v2 (Goyal et al., 2017b), Shapes VQA (Andreas et al., 2016), DocVQA (Mathew et al., 2021), OCR-VQA (Mishra et al., 2019), ST-VQA (Biten et al., 2019), Text-VQA (Singh et al., 2019), and GQA (Hudson and Manning, 2019).

Knowledgeable Visual Question Answering (KVQA) Unlike traditional VQA tasks focusing on the question relevant to the content image, KVQA requires the model to draw upon outside knowledge to answer questions. We incorporate two outside knowledge VQA datasets: OK-VQA (Marino et al., 2019) and A-OK-VQA (Schwenk et al., 2022), ScienceQA (Lu et al., 2022) which contains multi-modal science questions, and ViQuAE (Lerner et al., 2022) focusing on knowledge facts of named entities in images.

Classification This task involves classifying an image based on a given set of candidate labels. ImageNet (Russakovsky et al., 2015), Grounded Object Identification (COCO-GOI) (Lin et al., 2014), COCO-Text (Veit et al., 2016), Image Text Matching (COCO-ITM) (Lin et al., 2014), e-SNLI-VE (Kayser et al., 2021), Multi-modal Fact Checking (Mocheg) (Yao et al., 2022), and IQA (Duanmu et al., 2021) are included. Due to language model input length constraints, we reduce the number of options in some datasets with extensive candidate labels, such as ImageNet.

Generation Visual conditional general requires models to understand the visual content and make a composition meeting the task demand. We have Visual Storytelling (VIST) (Huang et al., 2016), Visual Dialog (VisDial) (Das et al., 2017), and multi-modal machine translation Multi30k (Elliott et al., 2016) in this category.

Chinese Vision-Language Tasks To examine the effect of instruction tuning on different languages, we incorporate several Chinese vision-language tasks including FM-IQA (Gao et al., 2015) for VQA, COCO-CN (Li et al., 2019) and Flickr8k-CN (Li et al., 2016) for captioning, Chinese Food Net (Chen et al., 2017) for classification, and MM-Chat (Zheng et al., 2022) for generation.

Video-Language Tasks Beyond the static images, we are interested in whether instruction tuning can be applied to video-language tasks. We include the classic MSR-VTT datasets (Xu et al., 2016) for video captioning, MSRVTT-QA (Xu et al., 2017), ActivityNet-QA (Yu et al., 2019), iVQA (Yang et al., 2021) and MSVD-QA (Xu et al., 2017) for video question answering, Something-Something (Goyal et al., 2017a) for video action classification.

Number of different instructions	400
- Image Captioning	52
- Classification	113
- Visual Question Answering	95
- Knowledgeable Visual QA	40
- Reasoning	60
- Generation	40
Tokens per instruction	24.4 ± 9.6
Instruction edit distance among the same task	76.6 ± 37.2
Instruction edit distance across tasks	106.6 ± 39.5

Table 2: The statistics of our instructions.

In summary, our dataset makes a wide coverage of the current existing visual-language and video-language benchmarks, enabling different skill sets for VLMs, from simple image captioning to complicated reasoning based on the image even beyond the visual content.

2.2 Annotation Process

To build high-quality multi-modal instruction datasets, we rewrite various datasets into a vision-to-text format. The annotation process includes three steps: (1) writing instructions for each task, (2) structuring images and texts into a unified schema, and (3) checking the overall dataset quality. Eight authors of this work are employed as human annotators, each of whom is a graduate student familiar with relevant literature.

Stage I: Instruction Writing To build high-quality instructions, we first ask annotators to carefully read the dataset paper and check the original dataset with some instances to get a clear understanding of the task. After that, they are required to write 10 diverse task instructions manually, covering the key characteristics of the task. Table 2 shows the statistics of the written instructions for each task and Figure 2 visualizes the instruction verb distribution. In total, we annotate 400 instructions for all tasks. The average length per instruction is 24.4. To evaluate the diversity of annotated instructions, we employ the average edit distance to measure the similarity between two strings. The average edit distance within the same task is 76.6, indicating a good range of instruction diversity.

Stage II: Data Format Unification After the instruction has been written according to the task

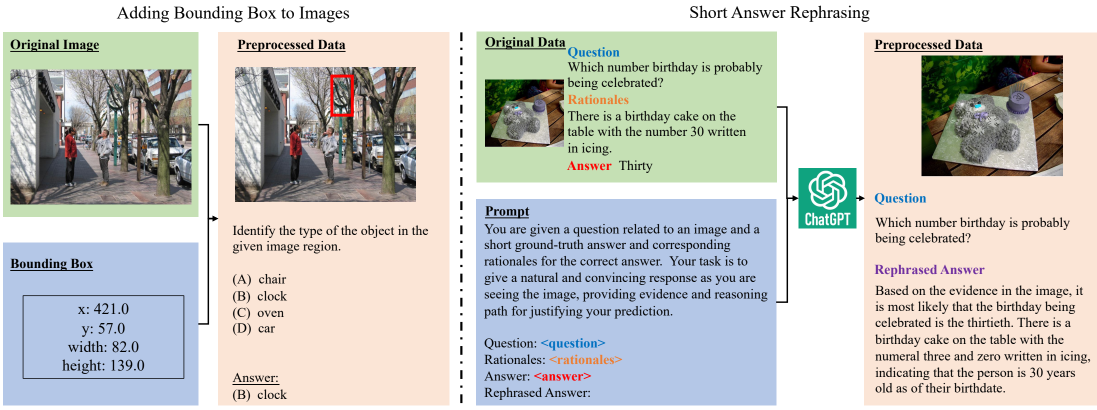


Figure 1: (Left) On region-based tasks, bounding boxes are added to serve as a visual referring prompt. (Right) Short answer rephrasing to improve the response quality, e.g., incorporating rationales into answers.

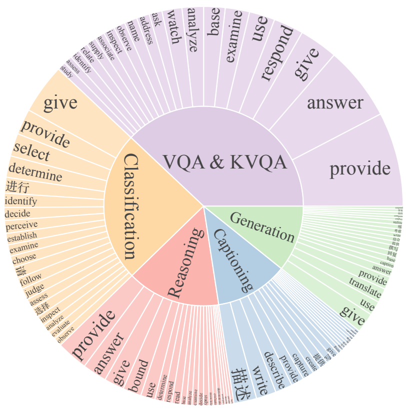


Figure 2: Top 20 verbs distribution for each task type.

Task	Description	Total #samples Train	Val	Test
CAP	Given an image, write a description for the image.	679,087	41,462	27,499
CLS	Given an image, classify the image into pre-defined categories.	238,303	100,069	21,206
VQA	Given an image, answer a question relevant to the image.	177,633	46,314	10,828
KVQA	Given an image, answer the question requires outside knowledge.	39,981	11,682	5,477
REA	Given an image, conduct reasoning over the images.	99,372	11,500	10,000
GEN	Given an image, make compositions with certain requirements.	145,000	11,315	17,350
Chinese	CAP, CLS, VQA, and GEN tasks in Chinese.	192,076	77,306	4,100
Video	CAP, CLS, and VQA tasks on video-language datasets.	20,868	7,542	9,294
Total	2,005,264			

Table 3: M2BIT task descriptions and statistics. We aggregate instance counts for training, validation, and test sets across all tasks, totaling 2M instances.

characteristics, we further process the images and corresponding text for a unified instance schema. For most datasets, we keep the original images and text, where images are converted into corresponding base64 encoded strings for easy data loading. We perform two modifications on potential examples: (1) **Adding Bounding Box to Images.** For tasks designed for specific regions in the image, a straightforward solution is to provide the bounding box information in natural language for informing the language models of the regions in interest. However, the image pre-processing techniques adopted by different vision encoders may resize the original image, and the original bounding box annotation thus needs further adjustments. Inspired by the recent observation that vision encoders such as CLIP (Radford et al., 2021) are sensitive to the visual prompt (Shtedritski et al., 2023), we directly tag the bounding box as a red rectangle to the image, serving as a visual referring prompting (OpenAI, 2023) for VLMs to focus on the target region. (2) **Short Answer Rephrasing.** As recent studies have shown that the original short and brief answers in the common VQA dataset could negatively influence the model generation performance (Dai et al., 2023; Chen et al., 2023b), we propose to utilize the ChatGPT (`gpt-3.5-turbo-0301`) (OpenAI, 2022) model for rephrasing the original answers, by providing origin question and answer with potential extra contextual information. Contextual information includes the caption of the original images, rationales for specific VQA tasks and OCR tokens for the scene-related question, which make the rephrased answers more engaging and informative. Figure 1 shows these two modifications.

Stage III: Quality Check In this stage, we assign a different annotator to each task to review 20 examples from each split of every dataset. During this stage, we identify minor format inconsistencies between tasks and address them by standardizing the task formats. We observe that answer rephrasing greatly improves the response quality, e.g., more than 95% instances we checked are perceived as better than original concise answers, while ChatGPT refused to rephrase a few answers (less than 3% of examined instances) due to insufficient image information. We employ simple heuristics to filter these failed answers and use a basic template to convert the original answer into a sentence. We find that this small portion of unsuccessful rephrased answers has negligible impact. Finally, the task dataset is deemed complete once the annotator can successfully load it and re-examine the accuracy of the instructions, inputs, and outputs for each instance examined.

2.3 Dataset Format and Statistics

The instance in our dataset consists of five fields: (1) **Images**: we represent the images with the potentially added bounding box by a base64 string. (2) **Instruction**: we randomly select an instruction from the task instruction pool for each instance. (3) **Inputs**: we allocate this field for providing task-specific inputs to the model, e.g., the question in the VQA tasks. For tasks such as captioning, there is no extra input so the corresponding field is left as an empty string. (4) **Outputs**: the required output to the specific tasks, such as the description of the image for captioning tasks and the answer to the image-related question. (5) **Meta Data**: we provide this field to preserve important information such as image ID for referencing the original dataset. With the clear distinction of these fields, the user of our benchmark can construct the training instances needed flexibly and evaluate the models conveniently. Table 3 gives the statistics aggregated by tasks, and we refer readers to Appendix B for a visualization of our schema, detailed statistics, and the license of each dataset.

3 Experiments

In this section, we built a VLM to verify the effect of M2BIT. We first introduce the experimental setups (§ 3.1), then report and discuss the evaluation results (§ 3.2). Lastly, we conduct an analysis on instruction tuning and provide a case study (§ 3.3).

3.1 Experimental Settings

Implementation Details Inspired by the recent success of BLIP-2 (Li et al., 2023a), we adopt the vision encoder and the Q-former architecture in the BLIP2-OPT-2.7B (Li et al., 2023a) model to extract relevant visual features from images. For the large language models, we utilize Ziya-13B (Zhang et al., 2022) derived from LLaMA (Touvron et al., 2023) with bilingual (English and Chinese) ability. We employ a two-staged training. **Stage I Visual-Text Alignment:** To align the visual and textual feature space, we utilize the instructions in the coco captioning and perform an initial alignment training on LAION 400M (Schuhmann et al., 2021). We train the Q-former and the language projection, resulting in a total 130M parameters to optimize with AdamW (Loshchilov and Hutter, 2019). The batch size is set to 256 to maximize the utilization of GPU and the model is trained with 300k steps. The learning rate linearly increases to a peak value of $5e$-5 in the first 2000 steps and follows a cosine decay scheduler. The weight decay is set to 0.05. **Stage II Multi-modal Instruction Tuning:** We further perform a multi-modal instruction tuning in our benchmark to activate the great potential of LLMs (see Figure 7 in Appendix B for used tasks). We train the model after alignment training for 3 epochs and with a lower learning rate of $1e$-5 and a warmup stage of 1000 steps. Inspired by LoRa tuning (Hu et al., 2022), the weights for mapping query and value vectors in the attention layer of LLMs are learnable in this stage to better adapt to the instruction tuning dataset. Other training parameters are consistent with Stage I. All experiments are conducted with 8 NVIDIA 80GB A100 GPUs. It took about 10 days for Stage I and Stage II can be finished in a day.

Evaluation Setup We conduct three evaluations to understand the instruction tuning effect comprehensively with our M2BIT: (1) Evaluation of KVQA tasks, which includes OK-VQA, A-OKVQA and a held-out dataset ViQuAE. These tasks are widely adopted in evaluation for VLMs (Dai et al., 2023; Bai et al., 2023) as they pose a great challenge for VLMs to understand the image context and perform reasoning with the knowledge acquired by LLMs. (2) Evaluation of object hallucination, which refers to a phenomenon that the model produces image descriptions that contain objects that are not anchored with or even absent from the target image. We follow the exact setup in Li et al. (2023b) to perform hallucination analysis on 2,000 images randomly sampled from the MSCOCO dataset (Lin et al., 2014). (3) Evaluation of cross-language/modality transferability. We hold out all Chinese V+L tasks and video-language tasks during the instruction tuning stage, then perform a zero-shot transfer to investigate whether instruction tuning is generalizable across languages, i.e., English to Chinese, and modalities, i.e., images to videos. In all experiments, we use greedy decoding in inference for deterministic results.

Metrics We adopt ROUGE-L (Lin, 2004) as an

Model	OK-VQA	A-OKVQA	ViQuAE
BLIP2-Flan-T5-XXL	9.1	15.6	9.7
MiniGPT4	23.3	21.8	24.4
InstructBLIP	7.1	5.9	7.3
Ying-VLM$_{\text{LLaVA}}$	26.4	22.5	24.3
Ying-VLM	**27.5**	**24.5**	**29.6**

Table 4: ROUGE-L evaluation results of KVQA tasks.

Model	Flickr-8k-CN	FM-IQA	Chinese-FoodNet
MiniGPT4	9.6	20.1	5.0
InstructBLIP	5.2	2.3	1.0
Ying-VLM (Ours)	**20.5**	33.3	**49.8**
+ trained w/ Flickr-8k-CN & FM-IQA	20.0	**39.8**	0.1

Table 5: Zero-shot transfer to Chinese vision-language tasks. Our model generalizes well on unseen Chinese captioning, VQA and classification tasks.

Model	Len	CHAIR$_I$(↓)	CHAIR$_S$(↓)	Avg. (↓)
mPLUG-Owl*	98.5	30.2	76.8	53.5
LLaVA*	90.7	18.8	62.7	40.8
MiniGPT-4*	116.2	9.2	31.5	20.4
InstructBLIP*	7.5	**2.5**	**3.4**	**3.0**
Ying-VLM$_{\text{LLaVA}}$	62.7	**11.0**	36.0	23.5
Ying-VLM	34.2	12.6	**16.8**	**14.7**

Table 6: Object hallucination evaluation with instruction "*Provide a brief description of the given image*". Len denotes the average length of generated captions. * denotes results collected from Li et al. (2023b).

automatic metric to assess the consistency between predictions and ground-truth answers, focusing on evaluating the model's conversational abilities. As the automatic metric may not fully capture the nuances of conversational quality, we further introduce GPT-4 as a proxy of human evaluators (§ 3.2). For the object hallucination, we follow Li et al. (2023b) to adopt CHAIR$_I$ and CHAIR$_S$ (Rohrbach et al., 2018). CHAIR$_I$ denotes the proportion of hallucinated ones in all generated objects, while CHAIR$_S$ describes the hallucination at the sentence level, i.e., the proportion of generated captions that contain hallucinated objects. Appendix F provides a detailed definition for these two metrics.

Baselines Recently proposed VLMs are adopted for comparison, including (1) BLIP-2-Flan-T5-XXL (Li et al., 2023a) where an instruction-tuned Flan-T5 (Wei et al., 2022) is connected with a powerful vision encoder to perform a series of multi-modal tasks; (2) MiniGPT-4 which aligns a CLIP visual encoder with a frozen Vicuna (Chiang et al., 2023) with artificially collected dialog dataset; (3) InstructBLIP, an instruction tuning enhanced VLM with Vicuna-13B with converted multi-model datasets and the LLaVA (Liu et al., 2023) dataset generated by text-only GPT-4. (4) Ying-VLM$_{\text{LLaVA}}$, we replace the M2BIT dataset with the LLaVA dataset (Liu et al., 2023) for instruction tuning with the same training setup, to isolate the effect of the base LLM.

3.2 Experimental Results

Evaluation of Knowledgeable VQA The results on the KVQA benchmarks are shown in Table 4. In comparison to the strongest baseline, our model achieves an improvement of 1.1 and 2.0 ROUGE-L points for OK-VQA and A-OKVQA, respectively. Additionally, Ying-VLM delivers the best performance on the held-out ViQuAE dataset. These findings indicate that instruction tuning on M2BIT effectively harnesses knowledge from LLMs and elevates response quality.

Evaluation of Object Hallucination As shown in Table 6, the VLMs trained on instruction tuning datasets generated by ChatGPT/GPT-4 exhibit serious hallucination problems, as indicated by the relatively high average CHAIR of LLaVA and MiniGPT-4. InstructBLIP performs the best on this evaluation. However, it should be noted that the too-short answers provided by InstructBLIP may result in a lack of politeness in responses, which can harm the user experience. This was validated in the later evaluation with GPT-4. When comparing Ying-VLM$_{\text{LLaVA}}$ to Ying-VLM, it can be observed that Ying-VLM achieves a much lower CHAIR$_S$ score (16.8 v.s. 36.0), demonstrating that it suffers significantly less from the sentence-level hallucination issue. These results suggest that M2BIT could help VLMs to achieve a better balance between the hallucination problem and response quality.

Cross-Language Transferability We assess models on three unseen Chinese vision-language tasks to investigate the cross-language generalization effect of instruction tuning. BLIP-2 and Flan-T5 are not considered here as they do not support Chinese outputs.[2] As illustrated in Table 5, our model performs well on all evaluated tasks compared with MiniGPT4 and InstructBLIP. While the gain can be attributed to the Chinese ability of the underlying Ziya-13B LLM, it promisingly indicates that instruction tuning with English datasets can effectively generalize to different languages. Further, we perform continual training on Chinese V+L tasks.

[2]For all models, we introduce a prompt to promote Chinese outputs. See Appendix D for details.

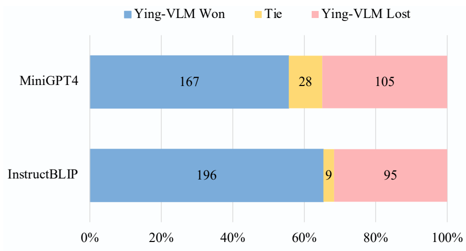


Figure 3: Evaluation results using GPT-4 as an evaluator. Our model outperforms MiniGPT-4 and InstructBLIP with a winning rate at 55.6% and 65.5%, respectively.

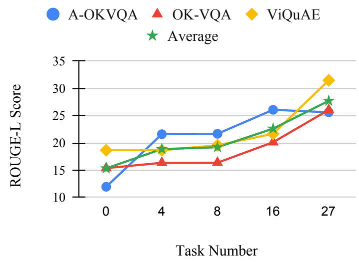


Figure 4: ROUGE-L score increases when models are trained with more instruction tuning datasets.

As shown in the last row of Table 5, the scores of FM-IQA can be further enhanced. However, the model suffers from catastrophic forgetting as there are no classification tasks in the continual training, resulting in poor performance on the Chinese-FoodNet. An in-depth investigation for this problem could be promising (Zhai et al., 2023).

Cross-Modality Transferability To evaluate performance on video-language tasks, we uniformly sample 8 frames from each video. MiniGPT4 is excluded as it does not support video inputs. Following InstructBLIP (Dai et al., 2023), we concatenate the visual embedding extracted from the Q-former of each frame as a prefix embedding to the language model. As demonstrated in Table 7, our model excels in these challenging settings, significantly surpassing the BLIP-series baselines. It is worth noting that the training dataset does not include any videos inputs, implying that our instruction tuning effectively aids the model in generalizing to inputs with a temporal dimension. Furthermore, continual training on video tasks in our dataset can improve the ROUGE-L scores on all tasks, indicating its effectiveness in boosting the video understanding abilities of VLMs.

GPT-4 Evaluation To further validate the quality of the generated response, we propose to utilize GPT-4 as a proxy of human evaluators (Peng et al., 2023; Gilardi et al., 2023). Specifically, following Vicuna (Chiang et al., 2023), we query GPT-4 to rate the performance of different models against our Ying-VLM. For each sample, we construct a prompt consisting of the original question, its corresponding reference answer, the response generated by our Ying-VLM, and a baseline system output. GPT-4 is asked to rate both responses on a scale of 10 based on the given question and its reference answer. The ratings are primarily based on the accuracy, relevance, and naturalness of the response to meet the requirements when humans are interacting with multi-modal agents (evaluation template is provided in Appendix E). We swap the order of candidate responses to mitigate potential evaluation biases (Wang et al., 2023). Considering the API cost of GPT-4, 300 examples are randomly sampled from OK-VQA, A-OKVQA and ViQuAE datasets as a subset for evaluation. Figure 3 shows that our Ying-VLM outperforms baseline models in most samples. For example, the GPT-4 evaluator favors Ying-VLM over MiniGPT4 on 167 over 300 tested samples. Consistent with the previous results, it indicates that VLMs tuned on M2BIT can produce more accurate and engaging responses on the challenging KVQA tasks.

3.3 Analysis

Effect of Task Number We investigate the influence of task numbers by randomly shuffling our tasks and then selecting a subset to train the model during the instruction tuning stage. Due to the computational resource limitation, we set up a maximum of 5k examples for each task and train all the models for 5k steps with a batch size of 64. We select 0, 4, 8, 16 and all 27 tasks for training, and report the individual ROUGE-L score and the average score. As illustrated in Figure 4, increasing the number of tasks greatly improves the results of the generalization performance. Besides, the performance gain is not diminished as the task number increases. This is promising as it indicates that we can continually improve performance by introducing more tasks into the training. Further analysis in Appendix G shows that the instruction diversity also affects the performance, and Ying-VLM performs robustly on unseen instructions.

Qualitative Results We conduct a case study to provide a more straightforward understanding of instruction-tuned models. The cases are chosen from the held-out ViQuAE and ChineseFoodNet datasets. As shown in Figure 5, Ying-VLM trained

Model	Video Captioning	Video Question Answering			
	MSRVTT	iVQA	ActivityNet-QA	MSRVTT-QA	MSVD-QA
BLIP-2-Flan-T5-XXL	8.8	11.1	8.9	10.3	13.2
InstructBLIP	14.3	6.3	9.3	4.0	7.0
Ying-VLM	14.2	23.5	21.9	18.3	21.4
+ trained w/ video datasets	**16.1**	**41.3**	**44.8**	**32.0**	**40.5**

Table 7: Zero-shot transfer to video-language tasks. We report ROUGE-L score for all tasks.

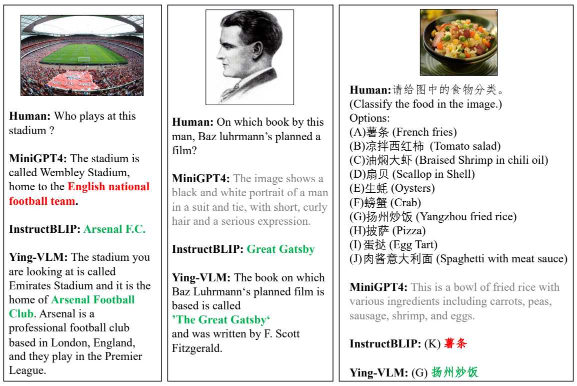


Figure 5: Case study of model outputs: Correct answers in green, wrong answers in red, and non-compliant answers in grey. Ying-VLM accurately responds to entity-centric questions and generalizes to unseen Chinese food classification tasks (English translation for visualization purposes only).

with M2BIT provides correct and informative responses to entity-centric questions requiring external world knowledge. In contrast, MiniGPT4 produces an incorrect answer for the stadium question on the left, and InstructBLIP consistently provides concise but less engaging answers. Ying-VLM also generalizes to Chinese inputs, accurately classifying the food image based on the instruction. These cases underscore the importance of instruction tuning dataset quality and demonstrate the effectiveness of the proposed M2BIT.

4 Related Work

Language and Multi-modal Instruction Tuning Language instruction tuning (Wei et al., 2022; Mishra et al., 2022) has been shown to enhance LLMs, enabling cross-task generalization (Longpre et al., 2023; Wang et al., 2022) and improved alignment with human intent (Ouyang et al., 2022). Recent research has expanded this concept to multi-modal instruction tuning for VLM development, evolving into two streams. The first uses established vision-text benchmarks to create an instruction-tuning dataset (Xu et al., 2022; Dai et al., 2023), while the second employs image annotation tools to generate a dialog-style dataset (Liu et al., 2023; Zhu et al., 2023; Zhao et al., 2023a).

Vision Language Models The success of LLMs has significantly propelled VLM development. The pioneering study Flamingo (Alayrac et al., 2022) and its open-source variants (Awadalla et al., 2023; Laurençon et al., 2023) have showcased the effectiveness of consolidating LLMs with vision encoders. PaLI-X (Chen et al., 2023c) delves deeper into the scaling effects of vision and language components. The Q-Former from BLIP-2 (Li et al., 2023a) has helped bridge the gap between the visual and text modalities. InstructBLIP (Dai et al., 2023) and MM-ICL (Zhao et al., 2023b) further integrate instructions into the visual-text alignment process for improved in-context learning ability (Dong et al., 2022). MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023) use a single projection layer, while mPLUG-Owl (Ye et al., 2023) adopts LoRA tuning (Hu et al., 2022), have shown promising results in aligning visual encoders and LLMs. The recently proposed Qwen-VL (Bai et al., 2023) has scaled up multi-modal pre-training and LLaVA-RLHF (Sun et al., 2023) explores RLHF (Ouyang et al., 2022) with LLaVA.

5 Conclusion

In this paper, we introduce M2BIT, a multi-modal bilingual instruction tuning dataset consisting of 2 million instances and 400 task instructions across 40 tasks. We develop Ying-VLM as a proof-of-concept model to demonstrate the effectiveness of our dataset. Compared with strong baselines, quantitative and qualitative results confirm that Ying-VLM outperforms KVQA tasks, exhibits reduced hallucination, and demonstrates superior generalization in unseen video and Chinese tasks. We anticipate that our proposed benchmark, pretrained models, and experimental findings will prove valuable for future research in the multi-modal domain.

Limitations

Limitations of Dataset Collection The number of Chinese tasks in our M2BIT is limited, as most high-quality multi-modal resources are English only. In the future, we look forward to incorporating more Chinese V+L tasks in our dataset and exploring machine translation techniques to improve the V+L task coverage in Chinese. Besides, our M2BIT focuses on image-to-text and video-to-text tasks, while more modalities, such as audio (Kim et al., 2019; You et al., 2022; Mei et al., 2023), can be considered further.

Limitations of Experimental Exploration In this paper, we curate the M2BIT dataset to provide a resource for developing powerful VLMs and explore the cross-lingual/modality effect of multi-modal instruction tuning. However, there are still underexplored setups worth investigating. Promising avenues include exploring improved methodologies for instruction and task selection by taking the interdependence of different tasks into consideration and exploring the effects of generalization across different languages and modalities. Furthermore, we only adopted the Ziya-13B LLM in our experiments due to its promising bilingual ability. Recently, many powerful foundation LLMs have been released, such as LLaMA-2 (Touvron et al., 2023) and Baichuan-2 (Baichuan, 2023). It would also be interesting to perform a comprehensive analysis regarding different model families and scales.

Ethic Considerations

In line with established practices in language instruction tuning (Mishra et al., 2022; Longpre et al., 2023), our M2BIT dataset has been carefully curated by gathering and unifying NLP datasets from various sources, including academic papers and projects, making them suitable for research purposes. The licenses for the included tasks can be found in Appendix B. However, it should be noted that there are certain tasks for which the license information is not publicly available. We strongly advise users to verify the license before using the dataset for non-academic purposes to avoid potential problems, and we would emphasize this in our released dataset project.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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. *arXiv preprint arXiv:2308.12966*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4290–4300.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023a. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023b. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel M. Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim M. Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023c. Pali-x: On scaling up a multilingual vision and language model. *ArXiv*, abs/2305.18565.

Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. 2017. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *ArXiv preprint*, abs/1705.02743.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning.

Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. 2021. Quantifying visual image quality: A bayesian view. *Annual Review of Vision Science*, 7:437–464.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

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. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2296–2304.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *ArXiv preprint*, abs/2303.15056.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017a. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. 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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

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*, pages 271–275.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv preprint*, abs/2301.13688.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4v(ision) system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *ArXiv preprint*, abs/2304.03277.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. 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 8748–8763.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *ArXiv preprint*, abs/2304.06712.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13878–13888.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

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

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *ArXiv preprint*, abs/1601.07140.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *ArXiv preprint*, abs/2212.10773.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1666–1677.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *ArXiv preprint*, abs/2205.12487.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality.

Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. 2022. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States. Association for Computational Linguistics.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *ArXiv preprint*, abs/2209.02970.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023b. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

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

A Datasheet for M2BIT

A.1 Motivation

For what purpose was the dataset created? M2BIT is created to facilitate multi-modal multilingual instruction tuning for large language models.

A.2 License

All annotated instructions are licensed under the CC-BY 4.0 license. For the licenses of original datasets, we refer users to Table 8 for more details.

A.3 Maintenance Plan

We commit to continually updating the dataset and rectifying any potential errors. Previous versions of the dataset can still be accessed in the Git history. Users can submit their questions and suggestions in the dataset hub, and we will promptly address their inquiries. We also encourage community contributions to expand the range of datasets by submitting pull requests to the dataset repository.

A.4 Composition

What do the instances that comprise the dataset represent? (e.g., documents, photos, people, countries) Our data is provided in JSON format.

Each data instance consists of (1) an instruction prompt, (2) a list of base64 strings representing images (3) a task-specific input, such as the question of the image, (4) a desired output, such as the answer for the image-related question, and (5) a metadata dictionary for referencing the original dataset.

How many instances are there in total (of each type, if appropriate)? The statistics of our dataset can be found in Table 8.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? We tried to transform the original whole dataset into a unified schema. However, due to the disk limitation and the cost of paraphrasing short answers, we chose a randomly sampled subset from the original dataset to perform the transformation.

Is there a label or target associated with each instance? Yes, the outputs field serves as the label.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? N/A.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes. We made the transformation based on the original dataset split.

Are there any errors, sources of noise, or redundancies in the dataset? No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Yes.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

A.5 Uses

Has the dataset been used for any tasks already? Yes. We have used the M2BIT dataset to train a vision-language model, which demonstrates promising results on knowledgeable VQA tasks and generalizes well to video-language tasks and Chinese vision-language tasks. Please see Section 4 of the main paper for details.

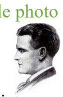
List[String]: the base64 string representation of a profile photo of F. Scott Fitzgerald

Images: ["iVBORw0KGg...5ErkJggg=="]
String: task instruction
Instruction: "Analyze the image and provide an appropriate response to the question. "
String: task-specific inputs, e.g., a question related to the image.
Inputs: "On which book by this man, Baz luhrmann's planned a film?"
String: task outputs, e.g., the correct answer for the question.
Outputs: "Baz Luhrmann has planned a film adaptation of the book The Great Gatsby. "
Dict: meta information dictionary contains original data.
Meta Data: {"kilt_id": "qw_1524", ... ,"wikipedia_id": "152171"}

Figure 6: Unified data format schema of our dataset.

What (other) tasks could the dataset be used for? M2BIT is a useful resource for instruction tuning studies in the multi-modal field. Future studies can utilize M2BIT to investigate the influence of instruction tuning and improve the general performance of vision-language models.

Is there a repository that links to any or all papers or systems that use the dataset? No.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The bounding boxes are added to the image with a red rectangle box to inform the model of regions in interest. For those models with a vision encoder that is not sensitive to these visual prompts, the effect of this operation can be limited. Besides, short answers in some VQA tasks are paraphrased by ChatGPT, which is designed to improve the response quality of the model while potentially impacting the language diversity of the model.

B Dataset Statistics

Table 8 lists the detailed statistics in our benchmark and Figure 6 illustrates the unified schema adopted in our dataset. We collect the dataset license from PaperWithCode.[3] For datasets under Unknown and Custom licenses, we suggest the users check the project page or contact the dataset owner before usage.

C Template for Answer Rephrasing

We provide the paraphrase template in Table 9 for querying the ChatGPT to re-write the original short

[3] https://paperswithcode.com/

Task	Dataset	Used	#samples			License
			Train	Val	Test	
Captioning	MS COCO (Lin et al., 2014)	Yes	566,747	25,010	25,010	Custom
	TextCaps (Sidorov et al., 2020)	Yes	97,765	13,965	0	Unknown
	Image-Paragraph-Captioning (Krause et al., 2017)	Yes	14,575	2,487	2,489	Custom
Classification	COCO-GOI (Lin et al., 2014)	Yes	30,000	2,000	0	Custom
	COCO-Text (Veit et al., 2016)	Yes	118,312	27,550	0	Custom
	ImageNet (Russakovsky et al., 2015)	Yes	30,000	50,000	0	Non-commercial
	COCO-ITM (Lin et al., 2014)	Yes	30,000	5,000	5,000	Custom
	e-SNLI-VE (Kayser et al., 2021)	Yes	20,000	14,339	14,740	Unknown
	Mocheg (Yao et al., 2022)	Yes	4,991	180	466	CC BY 4.0
	IQA (Duanmu et al., 2021)	Yes	5,000	1,000	1,000	Custom
VQA	VQA v2 (Goyal et al., 2017b)	Yes	30,000	30,000	0	CC-BY 4.0
	Shapes VQA (Andreas et al., 2016)	Yes	13,568	1,024	1,024	Unknown
	DocVQA (Mathew et al., 2021)	Yes	39,463	5,349	0	Unknown
	OCR-VQA (Mishra et al., 2019)	Yes	11,414	4,940	0	Unknown
	ST-VQA (Biten et al., 2019)	Yes	26,074	0	4,070	Unknown
	Text-VQA (Singh et al., 2019)	Yes	27,113	0	5,734	CC BY 4.0
	GQA (Hudson and Manning, 2019)	Yes	30,001	5,001	0	CC BY 4.0
KVQA	OK-VQA (Marino et al., 2019)	Yes	9,009	5,046	0	Unknown
	A-OK-VQA (Schwenk et al., 2022)	Yes	17,056	1,145	0	Unknown
	ScienceQA (Lu et al., 2022)	Yes	12,726	4,241	4,241	CC BY-NC-SA
	ViQuAE (Lerner et al., 2022)	No	1,190	1,250	1,236	CC By 4.0
Reasoning	CLEVR (Johnson et al., 2017)	Yes	30,000	2,000	0	CC BY 4.0
	NLVR (Suhr et al., 2017)	Yes	29,372	2,000	0	Unknown
	VCR (Zellers et al., 2019)	Yes	25,000	5,000	5,000	Custom
	VisualMRC (Tanaka et al., 2021)	Yes	15,000	2,500	5,000	Unknown
	Winoground (Thrush et al., 2022)	No	0	0	800	Unknown
Generation	Visual Storytelling (Huang et al., 2016)	Yes	5,000	4,315	4,350	Unknown
	Visual Dialog (Das et al., 2017)	Yes	50,000	1,000	1,000	CC By 4.0
	Multi30k (Elliott et al., 2016)	Yes	90,000	6,000	12,000	Non-commercial
Chinese	FM-IQA (Gao et al., 2015)	No	164,735	75,206	0	Unknown
	COCO-Caption CN (Li et al., 2019)	No	18,341	1,000	1,000	Non-commercial
	Flickr-8k-Caption CN (Li et al., 2016)	No	6,000	1,000	1,000	CC By 3.0
	Chinese Food Classification (Chen et al., 2017)	No	0	0	1,100	Unknown
	Multimodal Chat (Zheng et al., 2022)	No	3,000	1,000	1,000	Unknown
Video	Action-Classification (Goyal et al., 2017a)	No	2,000	2,000	2,000	Custom
	iVQA (Yang et al., 2021)	No	5,994	2,000	2,000	Unknown
	MSVD QA (Xu et al., 2017)	No	1,161	245	504	Unknown
	ActivityNet QA (Yu et al., 2019)	No	3,200	1,800	800	Unknown
	MSRVTT QA (Xu et al., 2017)	No	6,513	497	2,990	Unknown
	MSRVTT Captioning (Xu et al., 2016)	No	2,000	1,000	1,000	Unknown

Table 8: Detailed task descriptions and statistics of our instruction tuning tasks, including all datasets in all types of tasks. The column "Used" indicates whether we use this dataset in the instruction tuning stage.

answers, where {Q} and {A} is filled with the question and the answer need to be paraphrased, respectively. We incorporate an example to better inform the model of the paraphrasing tasks. For VQAv2 tasks, we add an extra {Caption} field in the template filled with corresponding captions from the COCO dataset to provide extra context information to help to rephrase. For A-OKVQA tasks, the rationale of each sample is adopted to enrich the final answer.

D Prompt for Zero-Shot Chinese Vision-Language Tasks

In our experiments, all VLMs are fine-tuned exclusively using English data. In our preliminary study, we observe that these models tend to generate English responses, even when the input and instructions are written in Chinese. We introduce a simple Chinese dialogue context during the zero-shot Chinese Vision-Language Task evaluation for all models, as illustrated in Table 10, Interestingly, this minor adjustment can encourage models to produce reasonable Chinese output. We leave the analysis of instruction-tuned VLM models' multi-

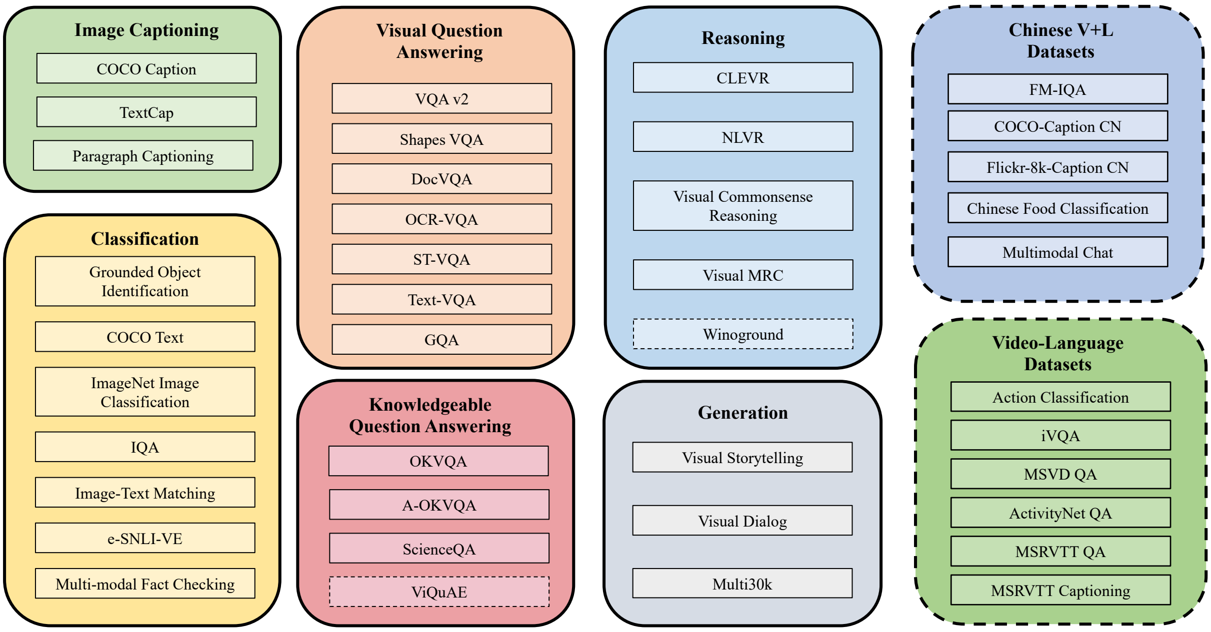


Figure 7: Tasks coverage of M2BIT. We hold out ViQuAE for evaluating the KVQA ability of VLMs and Winogound, as it only provides a small test set. All Chinese V+L and Video-language datasets are excluded during training for cross-lingual/modality effect investigation.

lingual capabilities for future research.

E Template for GPT-4 Evaluation

We adopt the template in Table 11 to query GPT-4 and obtain the evaluation results with FairEval [4] to obtain more stable results. Specifically, each tested instance is a quaternion: `(question, reference, response1, response2)`, where `response1` and `response2` are two responses from our Ying-VLM and the baseline model, respectively. For each instance, we query GPT-4 to judge which response is of better quality regarding accuracy, relevance and naturalness. We populate the quaternion into the evaluation template to form two query prompts: `T(Q=question, R=reference, R1=response1, R2=response2)` and `T(Q=question, R=reference, R1=response2, R2=response1)`. We set the temperature of GPT-4 to 1 and sample three completions for each query prompt. Therefore, each response will receive 6 scores, and we use the average score as the final score for each response. The response with the higher final score is considered the better response. The GPT-4 evaluation incurred a cost of \$20.45 for InstructBlip and \$20.90 for MiniGPT-4.

[4] https://github.com/i-Eval/FairEval

F Object Hallucination Metrics

For object hallucination evaluation, we adopt Caption Hallucination Assessment with Image Relevance (CHAIR) proposed by Rohrbach et al. (2018), a metric for evaluating object hallucination in image captioning tasks. Specifically, given the existing objects in the image, CHAIR calculates the proportion of objects that appear in the caption but not the image. CHAIR has two variants, i.e., CHAIR_I and CHAIR_S, which evaluate the hallucination degree at the object instance level and the sentence level, respectively. Formally, these two metrics are defined as:

$$\text{CHAIR}_I = \frac{|\{\text{ hallucinated objects }\}|}{|\{\text{ all mentioned objects }\}|}$$
$$\text{CHAIR}_S = \frac{|\{\text{ captions w/ hallucinated objects }\}|}{|\{\text{ all captions }\}|}.$$

Intuitively, CHAIR_I denotes the proportion of hallucinated ones in all generated objects, while CHAIR_S describes the hallucination at the sentence level, i.e., the proportion of generated captions that contain hallucinated objects. We follow the settings adopted in Rohrbach et al. (2018), which only consider 80 objects in the MSCOCO segmentation challenge. Following (Li et al., 2023b), a synonym list (Lu et al., 2018) is used for synonymous word unification in the generated captions to avoid misjudging hallucinated objects.

You are an AI visual assistant. Now you are given a question related to an image and a short ground-truth answer. Your task is to transform the ground-truth answer into a natural and convincing response. Make sure the response is accurate, highly relevant to the question, and consistent with the original answer.

Question:
Which NASA space probe was launched to this planet in 1989?
Answer:
Magellan
Transformed Answer:
NASA sent the Magellan spacecraft to Venus in 1989, which was the first planetary spacecraft launched from a space shuttle.

Question:
{Q}
Answer:
{A}
Transformed Answer:

Table 9: Template used to query ChatGPT for answer paraphrasing.

<human>:
请根据我的指示，以及所给的图片，做出相应的回答。
<bot>:
好的。
<human>:
{Instruction}
{Input}
<bot>:
好的。

Table 10: Prompt for promoting Chinese outputs.

G Effect of Instruction Diversity and Robustness

To investigate the influence of instruction diversity, we randomly select 1, 2, 4, and 8 instructions from each dataset, resulting in varied instruction diversity. The other training parameters are consistent with those used in previous experiments on task number investigation. Figure 8 shows that the performance varies with the level of diversity. Specifically, our results suggest that using four instructions per task is sufficient for achieving decent performance. We further explore the robustness of models on unseen instructions, where models are trained on 4 randomly selected instructions and evaluated with the left 6 instructions on each task. As shown in Figure 9, the model performs stably on the unseen instructions with a moderate 0.23 ROUGE-L score drop, indicating that it generalizes well on the unseen instructions.

[Question]
{Q}
[The Start of Reference Answer]
{R}
[The End of Reference Answer]
[The Start of Assistant 1's Answer]
{R1}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{R2}
[The End of Assistant 2's Answer]

[System]
We would like to request your feedback on the performance of two AI assistants in response to the user's multimodal question displayed above. We provided no multimodal inputs other than question text, but we provided a reference answer for this question. You need to evaluate the quality of the two responses based on the question and the reference answer.
Please rate the on the follow aspects:
1. Accuracy: whether the candidate's response is consistent with the original answer, this is important as we do not want a misleading result;
2. Relevance: whether the candidate's response is highly relevant to the question and image content;
3. Naturalness: whether the candidate's response is engaging, providing a great communication experience for the user when interacting with the AI visual assistant.
of the two Assistants' responses.

Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.
Then, output two lines indicating the scores for Assistant 1 and 2, respectively.

Output with the following format:
Evaluation evidence: <evaluation explanation here>
The score of Assistant 1: <score>
The score of Assistant 2: <score>

Table 11: Template used to query GPT-4 for evaluating the response quality of different models.

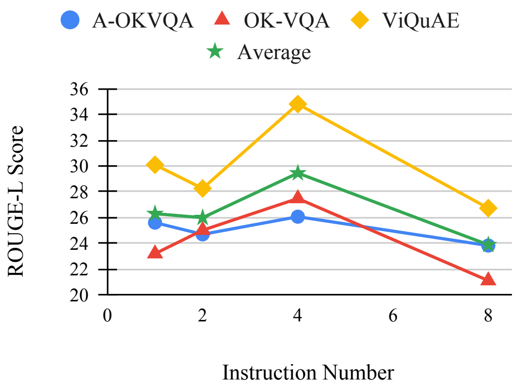


Figure 8: ROUGE-L Score changes with the varied number of instructions used for training.

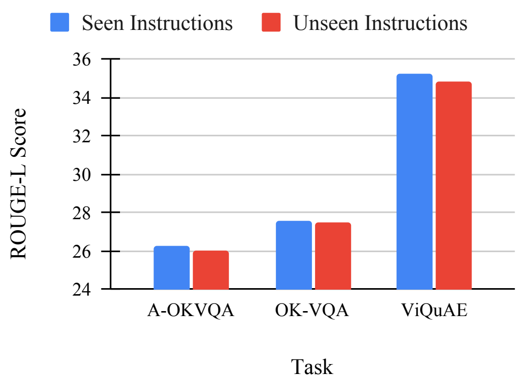


Figure 9: Ying-VLM performs stably on unseen instructions, with an average 0.23 ROUGE-L drop.