# EMPIRICAL STUDY TOWARDS BUILDING AN EFFECTIVE MULTI-MODAL LARGE LANGUAGE MODEL

**Skywork Multi-Modal Group**
Kunlun Inc.
willzhang@singularity-ai.com

## ABSTRACT

Despite substantial improvements in multi-modal tasks, especially in zero-shot settings, existing multi-modal large language models still suffer from several stubborn disadvantages. Current models occasionally produce inconsistent hallucinating outputs of the associated image. In addition, how to construct a bilingual (both English and Chinese) multi-modal large language model is still under-explored. In this paper, we examine what matters in data use, model design, and training pipeline, to reveal how to build a bilingual multi-modal large language model with fewer hallucinating outputs. Then, we present Skywork-MM, a multi-modal large language model, which consists of a frozen vision encoder, a learnable sampler module, and a frozen large language model tuned with a low-rank adapter (LoRA). We carefully build a bilingual instruction dataset to enhance the instruction following ability in both English and Chinese. Finally, we find it helpful to get a mental note first, and then answer specific questions as humans do. Through experimentation, we find that using significantly fewer training data (fewer than 50 million image-text pairs) allowed Skywork-MM to surpass all existing multi-modal large language models on the MME benchmark. Skywork-MM can also achieve competitive performance on MMBench and ScienceQA in the zero-shot setting. In order to assess the Chinese capability of current multi-modal language models, we conducted a qualitative analysis of typical Chinese scenes. Our code will be released at https://github.com/will-singularity/Skywork-MM.

## 1 INTRODUCTION

Large Language Models (LLMs) have demonstrated substantial capabilities in text generation and understanding, particularly in open-domain scenarios and zero-shot settings (Brown et al., 2020; Chung et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023). With the development of Supervised Instruction Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), LLMs further unlock impressive emergent capabilities that can solve large amounts of open-domain problems such as Math problems solving, question answering, and multi-round dialogue (Ouyang et al., 2022; Wang et al., 2022; Chiang et al., 2023; Peng et al., 2023). The use of large language models to address a variety of text-based tasks has been a resounding success, which has stimulated a growing interest in studying their integration with other modes of information. As a result, the field of multi-modal large language models has gained prominence and importance.

To this end, Flamingo (Alayrac et al., 2022) is first proposed to align a frozen visual encoder and a frozen large language model to achieve the in-context learning capability in the field of multi-modal tasks. BLIP2 (Li et al., 2023c) is a follow-up work that aligns vision features and LLMs with a QFormer bridge. Then, the instruction tuning method is migrated from using in text-only large language model to using in multi-modal large language model. A group of works such as LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), mPLUG-owl (Ye et al., 2023), InstructBLIP (Dai et al., 2023), and Lynx (Zeng et al., 2023) further improve the instruction-following ability of models by fine-tuning multi-modal instruction-following datasets. While these Multi-modal Large Language Models (MLLMs) demonstrate inspiring capabilities, they still suffer from several stubborn disadvantages. Existing MLLMs occasionally produce inconsistent hallucinating outputs of the associated image. In addition, how to construct a bilingual (both English and Chinese) multi-modal large language model with public data is still under-explored.


Figure 1: **Showing cases.** The top three images show that existing MLLMs sometimes fail to follow simple instructions and produce hallucinating outputs. The bottom three images show that current MLLMs fail to recognize typical Chinese scenes. Our method performs better in similar situations.

Existing MLLMs sometimes fail to follow the exact instructions or produce hallucinating outputs. For example, they prefer to answer “Yes" even when asking for the existence of an invisible object in the given image, as shown in Fig. 1. Our belief is that the shortcomings of these MLLMs lie in their limited exposure to a variety of training instructions, as they were trained to produce only existing (positive) visual concepts. However, it is also crucial to recognize and learn about the missing elements in the image, which were not explicitly instructed during the current training phase.

We also find that it is not natural to apply a bilingual language model to gain the ability of a bilingual MLLM like Chinese-LLaVA or ImageBind-LLM. Because cultural bias will exist in answering Chinese questions when we only use English corpus and vice versa. For example, in Fig. 1, current models, even fine-tuned with Chinese instruction data, fail to recognize items with typical Chinese characteristics.

To build a bilingual MLLM with fewer hallucinating outputs, we need to examine data usage, model design, and training pipeline and answer the following questions: 1) What data could be used and how to use it efficiently? 2) How to design the model to fit our data effectively? 3) What is the best

training pipeline in practice? In this paper, we propose a new MLLM, called Skywork-MM short for Skywork-MultiModal large language model, to address the above three questions.

**What data could be used and how to use it efficiently?** To address the problem of hallucinating outputs, we prepare to construct an image-central SFT dataset with well-labeled public datasets such as VQA (Antol et al., 2015), and OCRVQA (Mishra et al., 2019). Specifically, labeled images in various multi-modal tasks are from MSCOCO (Lin et al., 2014) and VG (Krishna et al., 2017). Therefore, we can construct different types of questions with one image in MSCOCO or VG, which will greatly enrich the diversity of instructions. In addition, we not only use questions about what can be seen in an image but also include questions about things that are absent in the image. Both positive and negative questions of the same image will be gathered in multiple rounds of dialog. This approach aims to equip the model with the ability to distinguish between what is present in the image and what is missing, thereby reducing the occurrence of erroneous outputs.

To build a strong bilingual MLLM, we must eliminate any cultural biases and enhance proficiency in following instructions, both in English and Chinese contexts. Therefore, we plan to use large-scale image-text pairs such as Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021), LAION (Schuhmann et al., 2021), and Taisu (Liu et al., 2022), which can provide a multi-cultural content integration. Although previous work (Zeng et al., 2023) noted that MLLMs do not benefit from large-scale but noisy image-text pairs because many of the texts in such datasets are not fluent or natural language expressions, it is still critical to use large-scale datasets to align vision concepts with language concepts, especially in removing cultural biases for bilingual MLLMs. To improve our ability to follow Chinese instructions, we use a method that involves translating English SFT data using GPT3.5. As a result, we generate a Chinese copy of the SFT data.

**How to design the model to fit our data effectively?** In order to address the issue of hallucinating outputs while building effective bilingual MLLM, several key model design considerations are necessary. First, to successfully align visual concepts with language concepts, we must ensure the appropriate usage of large-scale, low-quality web data such as LAION (Schuhmann et al., 2021; 2022) and Taisu (Liu et al., 2022). It is crucial to carefully eliminate any potential cultural biases when completing this step. Therefore, it is advisable to freeze the visual encoder to preserve the visual knowledge that has been learned through CLIP or other techniques, ensuring that this knowledge is not lost. Second, the output features of the visual encoder are somehow redundant. Therefore, we use a small learnable module like resampler in Flamingo (Alayrac et al., 2022), to extract useful vision features. The output of the resampler is a fixed number of query tokens which is dramatically less than the output tokens of the visual encoder. Third, as for the large language model in MLLM, there are usually two styles of tuning strategies. One is completely frozen like MiniGPT4 (Zhu et al., 2023) and InstructionBLIP (Dai et al., 2023), and the other is fine-tuning or adding adapter layers to large language models such as Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023b), mPLUG-owl (Ye et al., 2023), and LLaMA-Adapterv2 (Gao et al., 2023). In our line of work, we have incorporated a LoRA adapter (Hu et al., 2021), with the belief that it will enhance bilingual instruction-following capabilities and minimize hallucinating outputs by improving the alignment of a frozen visual encoder with large language models. Moreover, it is quite natural that the alignment of visual and text features should be carried out in a low-rank feature space.

**What is the best training pipeline in practice?** We split the training pipeline into two stages. For Stage 1, we use large-scale bilingual image-text pairs to learn the alignment of the visual encoder and the large language model. Note that the visual encoder is completely frozen, while the large language model is fine-tuned with LoRA adapter (Hu et al., 2021). For Stage 2, we build a unified chat prompt method to conduct supervised learning with our image-centric SFT data. In order to deal effectively with multi-modal tasks, we additionally propose a useful technique called mental notes. This technique works by describing the image in detail before moving on to solving the actual questions, which can significantly help achieve a resolution. Essentially, descriptions can be likened to the approach humans take when preparing mental notes to help them answer questions.

We empirically evaluate Skywork-MM on MME (Fu et al., 2023), MMBench (Contributors, 2023), ScienceQA (Lu et al., 2022), and found that, with significantly fewer training data (fewer than 50 million image-text pairs), Skywork-MM can still outperform existing English MLLMs in MME, and achieve competitive performance in MMBench and ScienceQA. In addition, we also conduct a qualitative analysis of everyday Chinese scenes, illustrating some examples of the country's exceptional multi-modal capabilities.

Our contributions are summarized as follows:

(1) We address the problem of how to build an effective bilingual multi-modal large language model with fewer hallucinating outputs in terms of data use, model design, and training pipeline.

(2) When engaging in a bilingual multi-modal large language model, it is of the utmost importance to remove cultural biases. In light of this, we propose a straightforward solution to enhance the multi-modal capabilities of the Chinese language.

(3) Introducing Skywork-MM, a multi-modal large language model that boasts exceptional capabilities and achieves state-of-the-art or competitive performance on MME. Our intention is to open access to our code and SFT data to stimulate the development of bilingual multi-modal large language models.

## 2 Related Work

**Large Language Models.** Large language models have shown great potential in solving a wide range of NLP tasks. A series of generative pre-trained transformers (GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), chatGPT, GPT-4 (OpenAI, 2023)) lead the trend of large language models and achieve very significant improvements in mathematics, reasoning, code, and other scenarios. In addition, other remarkable achievements such as PaLM (Chowdhery et al., 2022) with 540B learnable parameters demonstrate impressive natural language understanding and generation capabilities on several BIG-bench tasks (BIG-bench collaboration, 2021). LLaMa-2 (Touvron et al., 2023) with up to 70B learnable parameters also outperforms other open-source language models on many benchmarks, including reasoning, coding, proficiency, and knowledge tests. LLMs show surprising abilities, which are usually called emergent abilities in solving a series of complex tasks. It is convincing that LLMs can, to some extent, function as general-purpose language task solvers.

**Parameter-Efficient Fine-Tuning.** Adapter Tuning (Houlsby et al., 2019) is treated as a kind of parameter-efficient fine-tuning method for LLMs. It is usually achieved by serializing small neural network modules into Transformer models before the layer normalization block. In contrast to the serial adapter model, parallelizing parameter updates is another efficient fine-tuning implementation. Low-rank Adaptation (LoRA) (Hu et al., 2021) introduces a low-rank constraint to approximate the update matrix at each dense layer, thereby reducing the number of trainable parameters needed to adapt to downstream tasks. The basic concept behind LoRA is to freeze the original matrix and use low-rank decomposition matrices to approximate parameter updates. The most impressive advantage of LoRA is that it can significantly save memory and storage usage. Moreover, LoRA requires only a single copy of the original model, while maintaining multiple task-specific low-rank decomposition matrices for efficient adaptation to diverse downstream tasks. In particular, LoRA has been widely applied to open-source LLMs such as LLaMA (Touvron et al., 2023) and BLOOM (Workshop et al., 2023) for parameter-efficient fine-tuning.

**Large Vision Model.** Scaling models is a vital strategy for enhancing the quality of feature representation. In the field of computer vision, increasing the number of model parameters not only effectively bolsters the representation learning capacity of deep models, but also facilitates learning and knowledge acquisition from large datasets. Research on Large Vision Models (LVMs) has skyrocketed, exemplified by the success of models such as Vision Transformers (ViTs) (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021). In recent years, CLIP (Radford et al., 2021a) and ALIGN (Jia et al., 2021b), have shown a remarkable ability to extract visual features that are general in nature and are highly effective in various vision tasks. Eva-CLIP (Sun et al., 2023), a newly developed model, has enhanced the training techniques of CLIP, making it a more powerful and extensive vision model.

**Multi-Modal Large Language Models.** Multi-Modal Large Language Models (MLLMs) are designed to better mimic human perception of the world and often offer a more intuitive and user-friendly interface. They can support a larger spectrum of multi-modal tasks while LLMs can typically perform NLP tasks. MLLMs have emerged as a burgeoning research focus, employing potent LLMs as the cognitive powerhouse for executing diverse multi-modal tasks. Flamingo (Alayrac et al., 2022) makes use of a large language model in vision-language pre-training to solve the "in-context learning" problem for vision-language tasks. PaLI (Chen et al., 2022) jointly scales up the vision encoder and language encoder to cover a variety of language, vision, vision-language, and multilingual

tasks. Inspired by Flamingo, a series of MLLMs are proposed in the consideration of vision-language alignment methods (Li et al., 2023c; Dai et al., 2023; Ye et al., 2023; Zeng et al., 2023), and instruction tuning methods (Liu et al., 2023b; Zhu et al., 2023; Liu et al., 2023a). Unlike the above works, our focus is on efficiently constructing an effective MLLM that produces fewer hallucinating outputs, while also enabling robust bilingual multi-modal capabilities.

## 3 SKYWORK-MM

In this section, we will introduce the data usage, model design, and training pipeline of the proposed Skywork-MM.

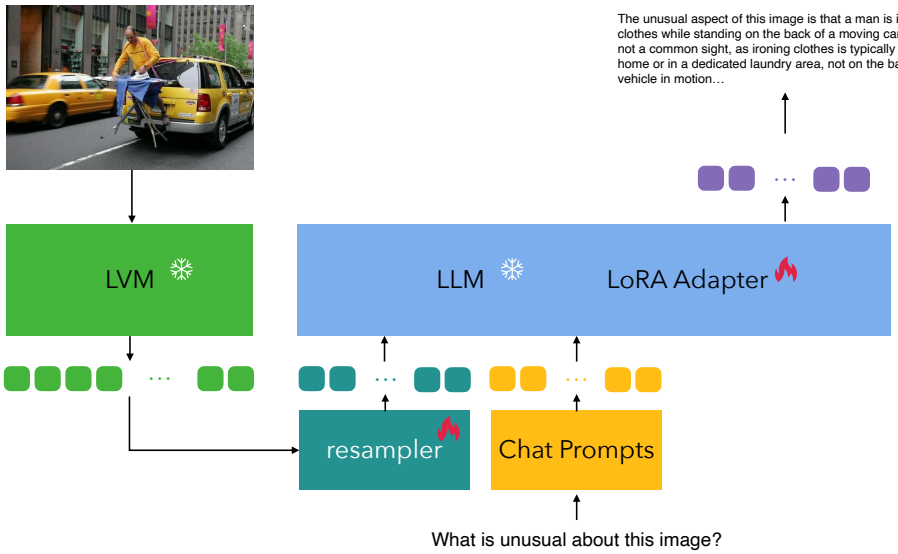


Figure 2: **The architecture and training pipeline of Skywork-MM.** Skywork-MM consists of four main modules, Large Visual Model (LVM), resampler, chat prompts, and Large Language Model (LLM). Given an image, we extract image features with LVM, and then the image features are fed into the resampler to compute tokens that can be accepted as LLM input. The LLM input is resampled tokens and instruction prompts (if any), and the output is image descriptions or response text (if any). Modules with a snowflake are frozen, and modules with a flame are trainable.

### 3.1 DATA USAGE

As noted in the introduction, we use two kinds of data, both of which are public data. One is large-scale image-text pairs from the Web, and the other is well-labeled Supervised Fine-Tuning (SFT) data. We will explain how to use both kinds of data in detail.

**Image-text pairs.** For English image-text pairs, we use Conceptual Caption (Changpinyo et al., 2021; Sharma et al., 2018), SBU (Ordonez et al., 2011) and LAION (Schuhmann et al., 2021; 2022). We use a subset in LAION called LAION-Aesthetics, which has been estimated by a model trained on top of clip embeddings to be aesthetic. Note that LAION-Aesthetics has a total of large data sizes, 120M. We only use a random subset of the dataset to balance distribution from different data sources. We find that it works well when the mixing ratio of CCSBU and LAION is 1:1. Therefore, the specific number of our used data can be found in Table. 1. Note that, Skywork-MM can obtain great performance even with much less data compared with existing MLLMs. For Chinese image-text pairs, we use an open-source dataset, TaiSu (Liu et al., 2022). We also sample a subset of TaiSu as shown in the Table. 1. Note that TaiSu data are only used in bilingual training, which produces Skywork-MM$_{bi}$.

| Data Type | Dataset | Image Domain | #Total | Public |
|---|---|---|---|---|
| Stage1 | SBU, CC-3M, CC-12M | Web | 14.9M | ✓ |
| Stage1 | LAION-Aesthetics | Web | 15M | ✓ |
| Stage1 | TaiSu (Liu et al., 2022) | Web | 10M | ✓ |
| Stage2 | Skywork-MM-EN | COCO, Flickr | 231k | ✓ |
| Stage2 | Skywork-MM-CN | Mixed | 15k | ✓ |
| Stage2 | OCRVQA (Mishra et al., 2019) | Web | 208k | ✓ |
| Stage2 | LLaVA (Liu et al., 2023b) | COCO, Flickr | 158k | ✓ |
| Stage2 | Minigpt4 (Zhu et al., 2023) | COCO, Flickr | 3.5k | ✓ |
| Stage2 | LRV (Liu et al., 2023a) | COCO, Flickr | 152k | ✓ |

Table 1: **Statistics of the datasets.** #Total denotes the total number of image-text pairs or image-instructions pairs.

**Supervised Fine-Tuning data.** We construct our SFT data, Skywork-MM-EN, with existing multi-modal tasks such as VQA (Antol et al., 2015), A-OKVQA (Schwenk et al., 2022) and OCRVQA (Mishra et al., 2019). Unlike the methods in InstructBLIP (Dai et al., 2023) and Lynx (Zeng et al., 2023), we organized the SFT data with an image-centric method, which means that we assign different questions for the same image, which can conduct a comprehensive understanding of the image. In addition, the questions posed to a given image encompass both positive and negative assertions, which will assist the model in generating less erroneous output. We also use existing multi-modal supervised fine-tuning data including LLaVA (Liu et al., 2023b), Minigpt4 (Zhu et al., 2023), and LRV (Liu et al., 2023a). Finally, we mix all the English SFT data and sample a subset to translate to Chinese with GPT3.5. We posit that a small amount of instruction data, when converted, has the potential to transfer the ability to follow instructions from English to Chinese. SFT samples are shown in Fig. 3.


{
"question": "Write a brief description for the photo."
"answer": "Three young children sit, peering intently at some subway tracks, a pile of luggage behind them."
},
{
"question": "For what group of people is the black area stepped on by the girl built for?"
"answer": "The group is blind people. The answer is blind"
},
{
"question": "How many children are sitting down?"
"answer": "3"
}

{
"question": "Can this truck protect items from weather in the pickup portion of the truck?"
"answer": "no"
},
{
"question": "What color are the stickers on the side of the truck?"
"answer": "red"
},
{
"question": "Are the men in uniform with the border control?"
"answer": "no"
}

{
"question": "为什么有人会把厕所放在这个不寻常的位置？"
"answer": "在公园或户外区域设立一个厕所，被草地和树木包围，如图所示，与典型的环境大相径庭。这可能是出于各种原因。例如，它可能是一个不寻常的艺术品或装置，旨在引发讨论、发表观点或引发对特定问题的思考。或者，它可能是某个需要临时设施的节日或活动的一部分，而厕所只是事件结束后被遗弃或错放的物品。最后，这也可能是一个恶作剧，有人故意将厕所放在那里以供娱乐或引起震惊。没有更多背景信息，很难确定确切的原因，但它肯定在场景中起到了非常不寻常和有趣的元素。"
}


Figure 3: **Sample of Supervised Fine-Tuned Data.**

### 3.2 Model Design and Training Pipeline

The Large Vision Model (LVM), the resampler, and the Large Language Model (LLM) are the three key modules we use. LVM offers high-quality vision features, but they are currently fixed in the

training pipeline and cannot be altered. We argue that the progress of LVM and MLLM should be decoupled, especially in training a bilingual MLLM with large-scale low-quality image-text pairs. The primary goal of LVM is to develop an understanding of visual concepts, while MLLM focuses on the alignment of visual and linguistic concepts. Therefore, we believe it is imperative that we prioritize freezing the LVM and focusing on learning alignment within the MLLM training pipeline. On the other hand, it is advisable to avoid tampering with LVM's parameters when aligning large-scale and low-quality image-text pairs. This may adversely affect the performance of understanding visual concepts. LVM is normally trained with contrastive learning such as CLIP (Radford et al., 2021b) and ALIGN (Jia et al., 2021a). In this work, we use ViT-g/14 in EVA-CLIP (Sun et al., 2023).

Resampler is borrowed from perceiver (Jaegle et al., 2021) and also used in Flamingo (Alayrac et al., 2022). We use a resampler to down-sampling visual features from a well-trained LVM. We incorporate LVM freezing into MLLM's complete training pipeline. A vital component that must be included is a small learnable module designed to enable the model to gain an understanding of the alignment between visual and linguistic concepts. In addition, the output of the resampler is a fixed number of query tokens which is dramatically less than the output tokens of the visual encoder which will boost the efficiency of the MLLM training pipeline.

In order to optimize the multi-modal large language model, we decided to add a LoRA adapter (Hu et al., 2021) while still maintaining the frozen parameters of the original model. This was done for the following reasons. When there are no learnable modules, the limitations of multi-modal capabilities and the inheritance of hallucination problems from large language models will become apparent. Conversely, choosing a more substantial adapter layer, such as LLaMA-adapter (Gao et al., 2023) and Lynx (Zeng et al., 2023), presents an obstacle in training, requiring large amounts of data, and in many cases text-only data. Hence, our belief is that the implementation of a LoRA adapter (Hu et al., 2021) will prove to be highly effective and efficient in facilitating the training of multi-modal large language models. Moreover, it is also a natural consequence that the alignment of visual and text features must occur in a low-rank dimensional feature space.

Similar to previous research (Ye et al., 2023; Zhu et al., 2023; Dai et al., 2023; Zeng et al., 2023), we have also implemented a two-stage training pipeline. Table 1 provides details of three image-text pair datasets used for Stage 1. These datasets are instrumental in teaching visual encoders and large language models how to align with each other. Moving on to Stage 2, our goal is to develop a unified approach to chat prompts for supervised learning using diverse datasets presented in Table 1. Our approach involves devising an image-central chat prompt that incorporates both positive and negative instructions, each related to an identical image. Positive instructions refer to questions that explore elements present in the image, while negative instructions refer to questions that explore elements absent from the image. Refer to Figure 3 for some examples. In order to deal effectively with multi-modal tasks, we have discovered a useful technique called mental notes. The technique used to solve the problem involves providing a detailed description of the image before attempting to answer the questions, which can significantly increase the likelihood of achieving a satisfactory resolution. By using a consolidated chat prompt for supervised fine-tuning, we can initiate the generation process by posing a query such as "Describe this image in detail." The model can then address specific questions. Mental notes can be considered a method that people use when preparing guidance notes to answer questions.

# 4 EXPERIMENT

## 4.1 IMPLEMENTATION DETAILS

We use ViT-g/14 from EVA-CLIP (Sun et al., 2023) as our LVM and LLaMA2-13B as our LLM. We use a resolution of 224*224 for images and LVM maps an image to a sequence of visual tokens with a patch size of 14. Following BLIP2 (Li et al., 2023c), we also remove the last layer of the ViT and use the second last layer's output features. The number of query tokens in the Resampler is 32, the hidden states width is the same as that of the large language model. We add a LoRA adapter to our LLM with $r = 64$ and $\alpha = 16$.

We pre-train for 120k steps in the first stage and 20k steps in the second stage. We use a batch size of 256 in the first stage and a batch size of 48 in the second stage. We use the AdamW (Loshchilov & Hutter, 2017) optimizer for both stages with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05.

Cosine learning rate decay is applied for both stages. The peak learning rate for Stage 1 is 1e-4, and the minimum learning rate is 8e-5. In Stage 2, the peak learning rate is 3e-5, and the minimum learning rate is 1e-6. During pre-training, we convert the frozen ViTs' and LLMs' parameters into FP16. Thanks to the use of frozen models and small datasets, our pre-training is very efficient. We use 400 GPU (Nvidia A100-80G) hours in total for Stage 1 and Stage 2, which means about 2 days with an 8-GPU-A100 machine.

### 4.2 Quantitative Evaluations

We first evaluate our method on two widely used multi-modal large language model benchmarks, MME (Fu et al., 2023) and MMBench (Contributors, 2023). We then evaluate our model on a classic held-out task, ScienceQA Lu et al. (2022), in the zero-shot setting. Incorporating a large language model and vision understanding capabilities into a system that can handle ScienceQA poses a significant challenge due to the complex and specialized knowledge required.

#### 4.2.1 MME

MME, a benchmark recently released, comprehensively evaluates large multi-modal language models by assessing their performance on 14 sub-tasks that span both perception and cognition tasks. MME covers the examination of perception and cognition abilities. Apart from OCR, the perception includes the recognition of coarse-grained and fine-grained objects. The former identifies the existence, count, position, and color of objects. The latter recognizes movie posters, celebrities, scenes, landmarks, and artworks. The cognition includes commonsense reasoning, numerical calculation, text translation, and code reasoning. MME evaluates a wide range of multi-modal capabilities and Skywork-MM can achieve the best average scores in comparisons with current MLLMs as shown in Table. 2, Table. 2 and Table. 4. We also ablate our models on MME. The results show that 1) mental notes are useful, especially on cognitive tasks, while without mental notes Skywork-MM can still significantly outperform other MLLMs on recognition tasks. 2) Although Skywork-MM$_{bi}$ exhibits an improvement in problem-solving capabilities in Chinese-centric scenarios (as demonstrated in Section 4.3), it simultaneously results in a slight decrease in the overall performance of its English-based multi-modal functionality. This phenomenon presents itself as a fascinating area of research, and we aim to investigate it in the future.

| Model | Existence | | Count | | Position | | Color | | OCR | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | |
| BLIP-2 (Li et al., 2023c) | 86.67 | 73.33 | 75.00 | 60.00 | **56.67** | 16.67 | 81.67 | 66.67 | 70.00 | 40.00 | 62.67 |
| LLaVA (Liu et al., 2023b) | 50.00 | 0.00 | 50.00 | 0.00 | 50.00 | 0.00 | 51.67 | 3.33 | 50.00 | 0.00 | 25.49 |
| MiniGPT-4 (Zhu et al., 2023) | 75.00 | 60.00 | 66.67 | 56.67 | **56.67** | **33.33** | 71.67 | 53.33 | 62.50 | 35.00 | 57.08 |
| mPLUG-Owl (Ye et al., 2023) | 73.33 | 46.67 | 50.00 | 0.00 | 50.00 | 0.00 | 51.67 | 3.33 | 55.00 | 10.00 | 34.00 |
| LLaMA-AdapterV2 (Gao et al., 2023) | 76.67 | 56.67 | 58.33 | 6.67 | 43.33 | 3.33 | 55.00 | 16.67 | 57.50 | 15.00 | 38.92 |
| InstructBLIP (Dai et al., 2023) | 95.00 | 90.00 | 80.00 | 63.33 | 53.33 | 13.33 | 83.33 | **70.00** | 57.50 | 15.00 | 62.08 |
| VisualGLM-6B (Du et al., 2022) | 61.67 | 23.33 | 50.00 | 0.00 | 48.33 | 0.00 | 51.67 | 3.33 | 42.50 | 0.00 | 28.08 |
| Otter (Li et al., 2023b) | 53.33 | 6.67 | 50.00 | 0.00 | 50.00 | 0.00 | 51.67 | 3.33 | 50.00 | 0.00 | 26.50 |
| Multimodal-GPT (Gong et al., 2023) | 46.67 | 10.00 | 51.67 | 6.67 | 45.00 | 13.33 | 55.00 | 13.33 | 57.50 | 25.00 | 32.42 |
| PandaGPT (Su et al., 2023) | 56.67 | 13.33 | 50.00 | 0.00 | 50.00 | 0.00 | 50.00 | 0.00 | 50.00 | 0.00 | 27.00 |
| Skywork-MM$^\star$ | 93.33 | 86.67 | 73.33 | 53.33 | 50.00 | 13.33 | **85.00** | **70.00** | 75.00 | 50.00 | 65.00 |
| Skywork-MM$_{bi}$ | **98.33** | **96.67** | 63.33 | 36.67 | 51.67 | 13.33 | 76.67 | 53.33 | **90.00** | **80.00** | 66.00 |
| Skywork-MM | 93.33 | 86.67 | **81.67** | **70.00** | 46.67 | 16.67 | 81.67 | 63.33 | 87.50 | 75.00 | **70.25** |

Table 2: **Evaluation of coarse-grained recognition and OCR.** In the MME dataset, there will be two questions for one image, whose answers are limited to "yes" or "no". Therefore, ACC is the accuracy calculated based on each question, and ACC+ indicates the accuracy based on each image where both of the questions need to be answered correctly. Avg. represents the average of all the numbers. All the reported numbers for the baseline methods are from Fu et al. (2023). Skywork-MM$^\star$ indicates that we evaluate our model without mental notes. Skywork-MM$_{bi}$ is the bilingual version model.

#### 4.2.2 MMBench

We also evaluate the comprehensive ability of our model in MMBench Liu et al. (2023c) of OpenCompass platform[1], which is a systematically-designed objective benchmark for robustly evaluating

[1] https://opencompass.org.cn/leaderboard-multimodal

| Model | Poster | | Celebrity | | Scene | | Landmark | | Artwork | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | |
| BLIP-2 (Li et al., 2023c) | 79.25 | 62.59 | 68.53 | 37.06 | 81.25 | 64.00 | 79.00 | 59.00 | 76.50 | **60.00** | 66.72 |
| LLaVA (Liu et al., 2023b) | 50.00 | 0.00 | 48.82 | 0.00 | 50.00 | 0.00 | 50.00 | 0.00 | 49.00 | 0.00 | 24.78 |
| MiniGPT-4 (Zhu et al., 2023) | 49.32 | 19.73 | 58.82 | 24.71 | 68.25 | 45.50 | 59.75 | 30.50 | 56.25 | 27.00 | 44.0 |
| mPLUG-Owl (Ye et al., 2023) | 77.89 | 57.14 | 66.18 | 34.12 | 78.00 | 57.50 | **86.25** | **73.00** | 63.25 | 33.00 | 62.63 |
| LLaMA-AdapterV2 (Gao et al., 2023) | 52.72 | 10.88 | 55.00 | 21.18 | 68.75 | 44.50 | 53.00 | 9.00 | 52.50 | 14.50 | 38.2 |
| InstructBLIP (Dai et al., 2023) | 74.15 | 49.66 | 67.06 | 34.12 | 84.00 | 69.00 | 59.75 | 20.00 | **76.75** | 57.50 | 59.20 |
| VisualGLM-6B (Du et al., 2022) | 54.42 | 12.24 | 50.88 | 2.35 | 81.75 | 64.50 | 59.75 | 24.00 | 55.75 | 20.00 | 42.56 |
| Otter (Li et al., 2023b) | 45.24 | 0.00 | 50.00 | 0.00 | 55.00 | 14.50 | 52.00 | 4.50 | 48.00 | 5.50 | 27.47 |
| Multimodal-GPT (Gong et al., 2023) | 45.24 | 17.01 | 49.12 | 24.12 | 50.50 | 17.50 | 50.50 | 23.00 | 46.00 | 12.00 | 33.50 |
| PandaGPT (Su et al., 2023) | 56.80 | 19.73 | 46.47 | 10.59 | 72.50 | 45.50 | 56.25 | 13.50 | 50.25 | 1.00 | 37.26 |
| Skywork-MM$^\star$ | 87.08 | 74.83 | **86.77** | **73.53** | 84.09 | 70.20 | 85.20 | 71.43 | 73.06 | 45.71 | **75.19** |
| Skywork-MM$_{bi}$ | 88.10 | 77.55 | 80.59 | 61.18 | **85.86** | **73.74** | 83.16 | 67.35 | 75.13 | 53.89 | 74.66 |
| Skywork-MM | **91.50** | **84.35** | **86.77** | **73.53** | 79.29 | 59.60 | 75.51 | 51.53 | 69.43 | 45.08 | 71.66 |

Table 3: **Evaluation of fine-grained recognition.** The settings are consistent with Table. 2. All the reported numbers for the baseline methods are from Fu et al. (2023).

| Model | Commonsense Reasoning | | Numerical Calculation | | Text Translation | | Code Reasoning | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | ACC | ACC+ | |
| BLIP-2 (Li et al., 2023c) | 68.57 | 41.43 | 40.00 | 0.00 | 55.00 | 10.00 | 55.00 | 20.00 | 36.25 |
| LLaVA (Liu et al., 2023b) | 49.29 | 11.43 | 50.00 | 0.00 | 52.50 | 5.00 | 50.00 | 0.00 | 27.28 |
| MiniGPT-4 (Zhu et al., 2023) | 58.57 | 34.29 | 47.50 | 20.00 | 42.50 | 15.00 | **67.50** | **45.00** | 41.30 |
| mPLUG-Owl (Ye et al., 2023) | 59.29 | 24.29 | 50.0 | 10.00 | 60.00 | 20.00 | 47.50 | 10.00 | 35.14 |
| LLaMA-AdapterV2 (Gao et al., 2023) | 54.29 | 14.29 | 52.50 | 5.00 | 52.50 | 5.00 | 52.50 | 10.00 | 30.76 |
| InstructBLIP (Dai et al., 2023) | **75.00** | **54.29** | 35.00 | 5.00 | 55.00 | 10.00 | 47.50 | 0.00 | 35.22 |
| VisualGLM-6B (Du et al., 2022) | 45.71 | 12.86 | 45.00 | 0.00 | 55.00 | 10.00 | 50.00 | 0.00 | 27.32 |
| Otter (Li et al., 2023b) | 48.57 | 10.00 | 47.50 | 10.00 | 55.00 | 10.00 | 50.00 | 0.00 | 28.88 |
| Multimodal-GPT (Gong et al., 2023) | 45.71 | 5.71 | 50.00 | 20.00 | 50.00 | 5.00 | 45.00 | 10.00 | 28.93 |
| PandaGPT (Su et al., 2023) | 56.43 | 17.14 | 50.00 | 0.00 | 52.50 | 5.00 | 47.50 | 0.00 | 28.67 |
| Skywork-MM$^\star$ | 64.29 | 34.29 | 55.00 | 15.00 | 50.00 | 0.00 | 50.00 | 15.00 | 35.45 |
| Skywork-MM$_{bi}$ | 66.43 | 37.14 | 52.50 | **30.00** | 45.00 | 5.00 | 45.00 | 0.00 | 35.13 |
| Skywork-MM | 72.14 | **54.29** | **65.00** | **30.00** | **60.00** | **20.00** | 45.00 | 10.00 | **44.55** |

Table 4: **Evaluation of cognition.** The settings are consistent with Table. 2. All the reported numbers for the baseline methods are from Fu et al. (2023).

the various abilities of vision-language models. The results for the dev set are shown in Table 5. We can see that Skywork-MM outperforms other MLLMs by a large margin. Note that Skywork-MM achieves remarkable success on attribute reasoning (AR) compared with competitors, over +5 points, implying that Skywork-MM can recognize characters and explore useful information in the image. Moreover, Skywork-MM consistently obtain strong performance on fine-grained perception (FP) task, indicating that Skywork-MM can effectively acquire detailed information in the visual scene.

#### 4.2.3 SCIENCEQA

Science Question Answering (ScienceQA) (Lu et al., 2022) is a challenging task that requires a model to utilize the information available across different modalities to synthesize a consistent and complete chain of thought. In ScienceQA, one's ability to peruse visual representations such as charts and tables is assessed. Thus, we evaluate our proposed approach in the zero-shot setting. It can be observed that Skywork-MM exhibits superior performance compared to other MLLMs that are based solely on the decoder-only large language model.

### 4.3 QUALITATIVE EXPERIMENTS

In addition to the examples depicted in Figure 1, we present an expanded selection of cases in Figure 4 to illustrate the versatility of our model in image recognition, instruction following, and the ability to understand Chinese characters. The first column contains four widely used test cases in MLLM. The second column shows examples with specific detailed instructions. The last column shows images from typical Chinese scenarios, which can not be solved by current MLLMs such as LLaVA and LLaVA-Chinese. Note that we use Skywork-MM$_{bi}$ to test Chinese cases. Qualitative evaluations indicate that not only does our model exhibit efficacy in conventional quantitative MLLM metrics such as MME and MMBench, but it also demonstrates remarkable expertise in understanding complex visual content, adhering to precise instructions, and identifying unique cultural characters.

| | Overall | LR | AR | RR | FP-S | FP-C | CP |
|---|---|---|---|---|---|---|---|
| JiuTian HITSZ (2023) | 69.0 | 48.3 | 71.6 | 73.0 | 68.4 | 57.9 | 79.9 |
| mPLUG-Owl (Ye et al., 2023) | 67.1 | 50.8 | 69.2 | 66.1 | 70.0 | 52.4 | 76.8 |
| MMICL Li et al. (2023a) | 67.9 | 49.2 | 71.6 | **73.0** | 66.7 | 57.2 | 77.2 |
| Shikra Chen et al. (2023) | 58.8 | 25.8 | 56.7 | 58.3 | 57.2 | 57.9 | 75.8 |
| Otter (Li et al., 2023b) | 51.4 | 32.5 | 56.7 | 53.9 | 46.8 | 38.6 | 65.4 |
| LLaVA (Liu et al., 2023b) | 38.7 | 16.7 | 48.3 | 30.4 | 45.5 | 32.4 | 40.6 |
| VisualGLM (Du et al., 2022) | 38.1 | 10.8 | 44.3 | 35.7 | 43.8 | 23.4 | 47.3 |
| MiniGPT-4 (Zhu et al., 2023) | 24.3 | 7.5 | 31.3 | 4.3 | 30.3 | 9.0 | 35.6 |
| Skywork-MM$^\star$ | 69.2 | 42.9 | 70.7 | 58.4 | 64.0 | 57.3 | 78.2 |
| Skywork-MM$_{bi}$ | 69.3 | **50.9** | 71.1 | 62.4 | 61.9 | 55.3 | 76.0 |
| Skywork-MM | **70.1** | 47.4 | **76.6** | 64.5 | **71.2** | **58.1** | **80.9** |

Table 5: **Evaluation of MMBench dev set.** All the reported performance for the baseline methods is from the leaderboard. Bold numers mean the best results for corresponding metrics. Skywork-MM$^\star$ indicates that we evaluate our model without mental notes. Skywork-MM$_{bi}$ is the bilingual version model.

| Decoder-only Models | | | Encoder-decoder Models | |
|---|---|---|---|---|
| BLIP-2 (Vicuna-13B) | InstructBLIP (Vicuna-13B) | Skywork-MM | BLIP-2 (FlanT5XXL) | InstructBLIP (FlanT5XXL) |
| 61.0 | 63.1 | **68.6** | 64.5 | **<u>70.6</u>** |

Table 6: **Evaluation of SciQA-Img at zero-shot settings** Bold numbers mean the best results for decoder-only models or encoder-decoder models and underlined number presents the overall best results.

## 5 Limitations and Conclusion

### 5.1 Limitations

Although we initially investigated the training of bilingual MLLM, the area of enhancing their multi-modal understanding of multiple languages is still under discussion. For example, as shown in Table. 2, Table. 4, and Table. 5, Skywork-MM$_{bi}$ is slightly weaker than Skywork-MM. It is of the utmost importance to eliminate all cultural biases and at the same time to gain in-depth knowledge of different cultural environments in order to overcome linguistic and cultural barriers. There are various potential solutions to the problem at hand. One option is to develop a multilingual LVM that can more effectively extract visual features specific to different cultures. Another option is to collect large-scale, high-quality image-text pairs in various languages. Ensuring the correct alignment of visual concepts with language concepts is of the utmost importance. We will devote future efforts to improving the current state of affairs.

In this paper, we only discuss how to build an effective MLLM based on a 13B model. It would be fascinating to examine larger language models. Strategies for refining parameters and training data requirements can vary significantly. There is also limited exploration of MLLM evaluation in the existing literature. The scope and instructions for MME and MMBench are still limited. We will leave these limitations to our future work.

It is still quite difficult for current MLLMs to accurately determine their positions, highlighting the ongoing challenge they face. In the MME benchmark, it appears that all existing MLLMs suffer from a relatively low ACC+ score in the position sub-task. It is likely that the reduced resolution and lack of positional descriptions in the training data contribute to this problem. However, it is critical for real-world applications, such as robotic perception, to overcome this obstacle. We will consider this as our next research effort.

### 5.2 Conclusion

In this paper, we first discuss how to build an effective multi-modal large language model with strong instruction-following ability and less hallucinating outputs from the perspective of data usage, model design, and training pipeline. We then train a new model, called Skywork-MM, that can achieve state-of-the-art performance on two widely used MLLM benchmarks. We also reveal that cultural biases are crucial to building an effective bilingual MLLM, and then propose a simple fix to train an effective bilingual Skywork-MM, which works better in typical Chinese scenes.


Figure 4: **Qualitative results.** The first column details common test cases used to evaluate MLLM performance. The second column outlines cases with specific instructions. The third column shows MLLM's multifaceted capabilities in representative Chinese scenarios.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.

BIG-bench collaboration. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*, 2021. URL https://github.com/google/BIG-bench/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311, 2022. URL https://arxiv.org/abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/InternLM/OpenCompass, 2023.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

HITSZ. Jiutian, 2023. URL https://github.com/rshaojimmy/JiuTian.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *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*, pp. 4904–4916. PMLR, 2021a. URL http://proceedings.mlr.press/v139/jia21b.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *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*, pp. 4904–4916. PMLR, 2021b. URL http://proceedings.mlr.press/v139/jia21b.html.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

OpenAI. Gpt-4 technical report, 2023.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1143–1151, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

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

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *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*, pp. 8748–8763. PMLR, 2021a. URL http://proceedings.mlr.press/v139/radford21a.html.

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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *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*, pp. 8748–8763. PMLR, 2021b. URL http://proceedings.mlr.press/v139/radford21a.html.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max

Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023.

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