• RESEARCH PAPER •

MMInstruct: A High-Quality Multi-Modal Instruction Tuning Dataset with Extensive Diversity

Yangzhou LIU[4†], Yue CAO[4†], Zhangwei GAO[7,1†], Weiyun WANG[5,1†], Zhe CHEN[4,1†], Wenhai WANG[6,1†], Hao TIAN[2], Lewei LU[2], Xizhou ZHU[3,1,2], Tong LU[4], Yu QIAO[1] & Jifeng DAI[3,1*]

[1]*Shanghai AI Laboratory, Shanghai* 200232*, China;*
[2]*SenseTime Research, Shanghai* 200233*, China;*
[3]*Tsinghua University, Beijing* 100084*, China;*
[4]*Nanjing University, Nanjing* 210023*, China;*
[5]*Fudan University, Shanghai* 200433*, China;*
[6]*The Chinese University of Hong Kong, Hong Kong* 999077*, China;*
[7]*Shanghai Jiao Tong University, Shanghai* 200240*, China*

Abstract Despite the effectiveness of vision-language supervised fine-tuning in enhancing the performance of Vision Large Language Models (VLLMs). However, existing visual instruction tuning datasets include the following limitations: (1) Instruction annotation quality: despite existing VLLMs exhibiting strong performance, instructions generated by those advanced VLLMs may still suffer from inaccuracies, such as hallucinations. (2) Instructions and image diversity: the limited range of instruction types and the lack of diversity in image data may impact the model's ability to generate diversified and closer to real-world scenarios outputs. To address these challenges, we construct a high-quality, diverse visual instruction tuning dataset MMInstruct, which consists of 973K instructions from 24 domains. There are four instruction types: Judgement, Multiple-Choice, Long Visual Question Answering and Short Visual Question Answering. To construct MMInstruct, we propose an instruction generation data engine that leverages GPT-4V, GPT-3.5, and manual correction. Our instruction generation engine enables semi-automatic, low-cost, and multi-domain instruction generation at 1/6 the cost of manual construction. Through extensive experiment validation and ablation experiments, we demonstrate that MMInstruct could significantly improve the performance of VLLMs, e.g., the model fine-tuning on MMInstruct achieves new state-of-the-art performance on 10 out of 12 benchmarks. The code and data shall be available at https://github.com/yuecao0119/MMInstruct.

Keywords instruction tuning, multi-modal, multi-domain, dataset, vision large language model

1 Introduction

Benefiting from the large-scale parameters and extensive pre-training corpus, Large Language Models (LLMs) [14, 49, 63, 65, 66] have demonstrated a range of powerful capabilities, including language generation, in-context learning, world knowledge, and commonsense reasoning. Beyond the pre-training phase, these models undergo an additional training stage, termed instruction tuning, which equips these base models with the ability to follow user instructions, thus transforming them into chat models. By integrating these chat models with pre-trained vision foundation models through a vision-language connector, Vision Large Language Models (VLLMs) exhibit impressive performance across various vision-language tasks. These models employ similar training schemes to empower VLLMs to effectively understand visual information. Specifically, during the pre-training phase, models are trained to predict the next text token conditioned on a given image, while during the instruction tuning stage, the models are required to learn to interact with users conditioned on the given image and instructions.

However, existing multi-modal instruction tuning datasets [10, 40, 68] include following limitations: (1) **Image Diversity**: The images of these datasets are sourced from existing datasets, such as COCO [37],

† Equal contribution.
* Corresponding author (email: daijifeng@tsinghua.edu.cn)

Table 1 Comparison of MMInstruct with existing visual instruction tuning dataset. Note that we unify the division of instruction tasks for all datasets based on our task domain partitioning. Question types are abbreviated due to space constraints. TF: judgment; MC: multiple-choice; LVQA: Long VQA; SVQA: Short VQA.

Dataset	#Instances	#Domains	Question Types	Question Form
LLaVA [40]	150K	3	LVQA, SVQA	Fixed
ShareGPT4V [10]	100K	1	LVQA	Fixed
M^3IT [34]	2.4M	12	TF, MC, LVQA	Fixed
Shikra [9]	156K	10	LVQA	Diverse
InstructBLIP [15]	1.6M	12	TF, MC, LVQA	Fixed
MultiInstruct [75]	510K	14	TF, MC, LVQA	Diverse
Vision-Flan [74]	1.6M	22	TF, MC, LVQA	Fixed
MMInstruct (Ours)	973K	24	TF, MC, LVQA, SVQA	Diverse

which is restricted to the common scenes and thus limits the models' generalization ability. For instance, models struggle to process the text-oriented OCR image. (2) **Annotation Quality**: These datasets are generated automatically by employing models (*e.g.*, GPT-4V [51]) to generate new question-answer pairs based on annotations from existing datasets. Despite the advanced capabilities of existing VLLMs, such data generation pipelines inevitably introduce noise to the generated dataset, leading to hallucinations in models. (3) **Instruction Diversity**: The instruction types within these datasets are limited, negatively impacting the models' ability to generalize across the diverse range of real-world instructions.

To address these issues, we propose a high-quality and diverse visual instruction tuning dataset named MMInstruct, which contains 973K instructions. To achieve the universality of the dataset, we design 24 task domains commonly seen in daily life, including (1) Perception (image style, image scene, image quality, image comparison, object localization, object relation, attribute recognition, image description, OCR, posters, artwork, landmark, spatial relationship, brand recognition, species recognition); (2) Reasoning (numerical calculation, image emotion, commonsense reasoning, complex reasoning, social relation, future prediction, meme comprehension, writing); (3) Multi-Round Long Visual Question Answering (Multi-Round Long VQA). We show some example instructions of various question types in Figure 2 and different domains in Figure 3. Specifically, our instructions comprise four common types: Judgement, Multiple-Choice, Long Visual Question Answering (Long VQA), and Short Visual Question Answering (Short VQA). The instructions do not adhere to a fixed template, and there may be variations in format among instructions with the same questioning purpose. And we additionally construct Multi-Round Long VQA data for long-context logical reasoning training of the model.

Relying on manual efforts to construct such a diverse and high-quality dataset can be excessively expensive, especially when the data scale is large. Therefore, we propose a semi-automatic, low-cost instruction generation data engine that leverages GPT-4V [51], GPT-3.5 [49], and manual correction. To enrich the scope of image coverage, we first utilize web crawlers and similarity searches to swiftly gather a large quantity of high-quality images. Then, these images undergo deep semantic analysis via GPT-4V, transcending the mere reliance on rudimentary annotations of the images themselves. After that, we integrate the characteristics of both the images and domains to design approximately ten seed questions for each domain. Unlike other datasets, the seed questions in our engine serve merely as references, encouraging GPT to generate diverse forms of instructions. Specifically, questions and answers are generated at the same time to ensure accuracy and reduce illusions. In this way, the data engine can automatically generate detailed semantic captions and diverse instructions for the image. Additionally, manual corrections are integral throughout the entire process to ensure dataset quality and minimize biases.

As shown in Table 1, we compare MMInstruct with some representative visual instruction tuning datasets, demonstrating significant advantages in terms of coverage and diversity of our instructions. Furthermore, when compared to the exclusive dependence on manual dataset construction, our data engine's cost is only 1/6 of manual annotation while concurrently ensuring data quality. The cost comparison between manual construction and MMInstruct is shown in Table 2.

Table 2 Comparison of costs between MMInstruct construction and manual construction. **Total** refers to the estimated cost of building the MMInstruct.

Method	Manual Construction	MMInstruct
Per Image	-	\$0.00885
Per Instruction	\$0.84	\$0.0004
Total	\$817, 320	\$128, 304

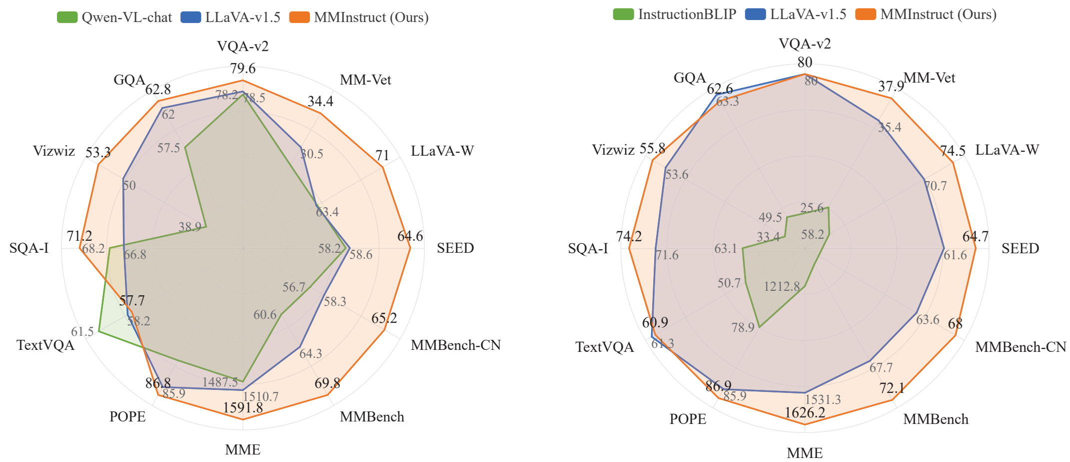


Figure 1 Performance comparison of different model sizes. (a) Compared with 7B models including Qwen-VL-Chat [2], LLaVA-1.5-7B [40], our model achieves SoTA on 11 benchmarks. (b) Compared with 13B models, including InstructBLIP [15], LLaVA-1.5-13B [40], our model achieves SoTA on 10 benchmarks.

To verify the effectiveness of MMInstruct, we incorporate MMInstruct into the instruction fine-tuning phase of LLaVA-1.5 [40]. Our experimental results demonstrate that MMInstruct significantly enhances the capabilities of VLLMs. Figure 1 shows the performance comparison of LLaVA-1.5 [40] on different benchmarks after fine-tuning on LLaVA-665K and MMInstruct. We can see that after fine-tuning on MMInstruct, our model demonstrates impressive improvements across a wide range of evaluation benchmarks and exceeds LLaVA-1.5 on 10 out of 12 benchmarks. We also conduct extensive ablation experiments to analyze the impacts of varying the fine-tuning data on VLLMs. These results highlight the effectiveness of MMInstruct.

In conclusion, our paper makes the following contributions:

• We construct a visual instruction tuning dataset MMInstruct, containing 24 common domains. MMInstruct comprises 973K high-quality and diverse visual instructions featuring diverse question forms and types, including judgment, multiple-choice, Long VQA, and Short VQA.

• To construct MMInstruct, we designed a semi-automatic, low-cost instruction generation data engine based on GPT-4V, GPT-3.5, and manual correction. Compared with purely manual construction, Our data engine's cost is only 1/6 of manual annotation while ensuring annotation quality and data diversity.

• We conduct comprehension experiments to validate the effectiveness of MMInstruct. As shown in Figure 1, after fine-tuning on MMInstruct, LLaVA-1.5 achieves state-of-the-art results on 10 out of 12 benchmarks. Specifically, the scores on MME [18] and LLaVA-Bench (In-the-Wild) [41] are 1626.2 and 74.5, surpassing LLaVA-1.5 by 94.9 and 3.8 points respectively.

2 Related Work

Vision Large Language Models. Significant progress has been achieved in the realm of Vision Large Language Models (VLLMs). Models like CLIP [54], ALIGN [23], EVA [17], which are trained via contrastive learning-based methods, demonstrate the capacity to understand the complex semantics of the open-world through image-text alignment. Subsequent endeavors, as exemplified by VL-BERT [60], VL-BEiT [4], ALBEF [32], VLMo [3], BEiT-3 [70], CoCa [78], and the Uni-Perceiver series [30, 83, 84], have shown proficiency in performing a variety of multi-modal downstream tasks. However, these models are trained from scratch, leading to escalated expenses in the development of novel models.

In recent years, numerous VLLMs [1, 11–13, 16, 27, 43, 64, 69, 73] have been developed by incorporating pre-trained LLMs [14, 63, 65, 66] with off-the-shelf vision encoders (*e.g.*, CLIP [54]), aiming to combine the visual encoding capabilities of vision encoders alongside the linguistic knowledge of language models. Earlier research, such as Frozen [67] and VisualGPT [8], demonstrates the efficiency of employing LLMs as decoders for VLLMs, facilitating learning from multi-modal data. Flamingo [1] can employ interleaved texts and images as input and is endowed with remarkable few-shot learning capabilities. In VLLMs, a feature resampler [31] or projection layer [40, 41] is employed to align the features of vision encoders

Question: Is the character in the image "TEXACO"?
Choices: Yes / No

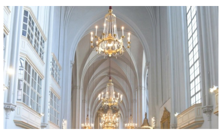
Question: Are the children playing football on the football field?
Choices: Yes / No

(a) Judgement

Question: Which movie is this picture from?
Choices:
A. The Avengers
B. Iron Man
C. Spider-Man
D. Black Widow

Question: What does this picture make you feel?
Choices:
A. Boredom and Ennui
B. Curiosity and fun
C. Happy and excited
D. Fierce and fearful

(b) Multiple-Choice



Question: What is the name of the cathedral in this picture?
Answer: St. Augustine Church in Vienna.

Question: What is the dominant style of the cat in this picture?
Answer: Realistic style.

(c) Short VQA

Question: What is the name of the state marked "AK" on the map?
Answer: The state labeled "AK" is Alaska. We can clearly see that the state labeled "AK" is in the lower left corner, and according to the convention of common maps, "AK" on a map of the United States represents Alaska.

Question: What is the relationship between cyclists?
Answer: The cyclists are depicted in different colors, showing their determination and focus. They are chasing each other in the race, showing a strong sense of competition and desire to win.

(d) Long VQA

Caption: This is the promotional poster for the Disney Pixar animated film "Coco". At the top of the poster, the "DISNEY PIXAR" logo prominently indicates that this is a collaboration between Disney and Pixar. In the center, the title of the movie "COCO" is displayed in colorful fonts, with some letters partially obscured, showing as "CO" and "TR", which may be a creative element in the poster design. Above the letters are decorative star shapes and vibrant colors, which serve to attract attention. Below the title, it reads "MUY PRONTO SÓLO EN CINES" in Spanish, translated to English as "Coming Soon Only in Theaters", informing the audience that the movie will be released in theaters soon...

(e) Image Caption

Figure 2 Examples of various question types in MMInstruct. (a) to (e) represent different question types in MMInstruct. **Question** denotes instruction generated by GPT. **Answer** denotes the response based on the instruction. **Caption** denotes the detailed image description generated by GPT. The green option indicates the correct answer.

with the embedding space of language modes. This alignment facilitates the LLM in acquiring the ability to understand images. With the introduction of visual instruction tuning in VLLMs (*e.g.*, InstructBLIP [15], Qwen-VL [2], InternVL [12,13], GPT-4 [51], LLaVA series [40,41], Gemini series [55,62]), a significant enhancement has been observed in the capability to follow instructions and complete visual tasks. However, many advanced models [2,51] do not publish their SFT datasets. Currently, the community urgently needs a diverse, high-quality, open-source visual instruction dataset to further improve model performance.

Datasets for Vision-Language Supervised Fine-tuning. In the NLP community, the utilization of instruction-following data [53,71,72] during the SFT stage enables Large Language Models (LLMs) to acquire the capability of following natural language instructions and solving real-world tasks, thus contributing to notable advancements [14,50,52,66]. The integration of the vision modality further enhances this process by providing additional information for interactions, making visual instruction tuning a more creative and innovative procedure.

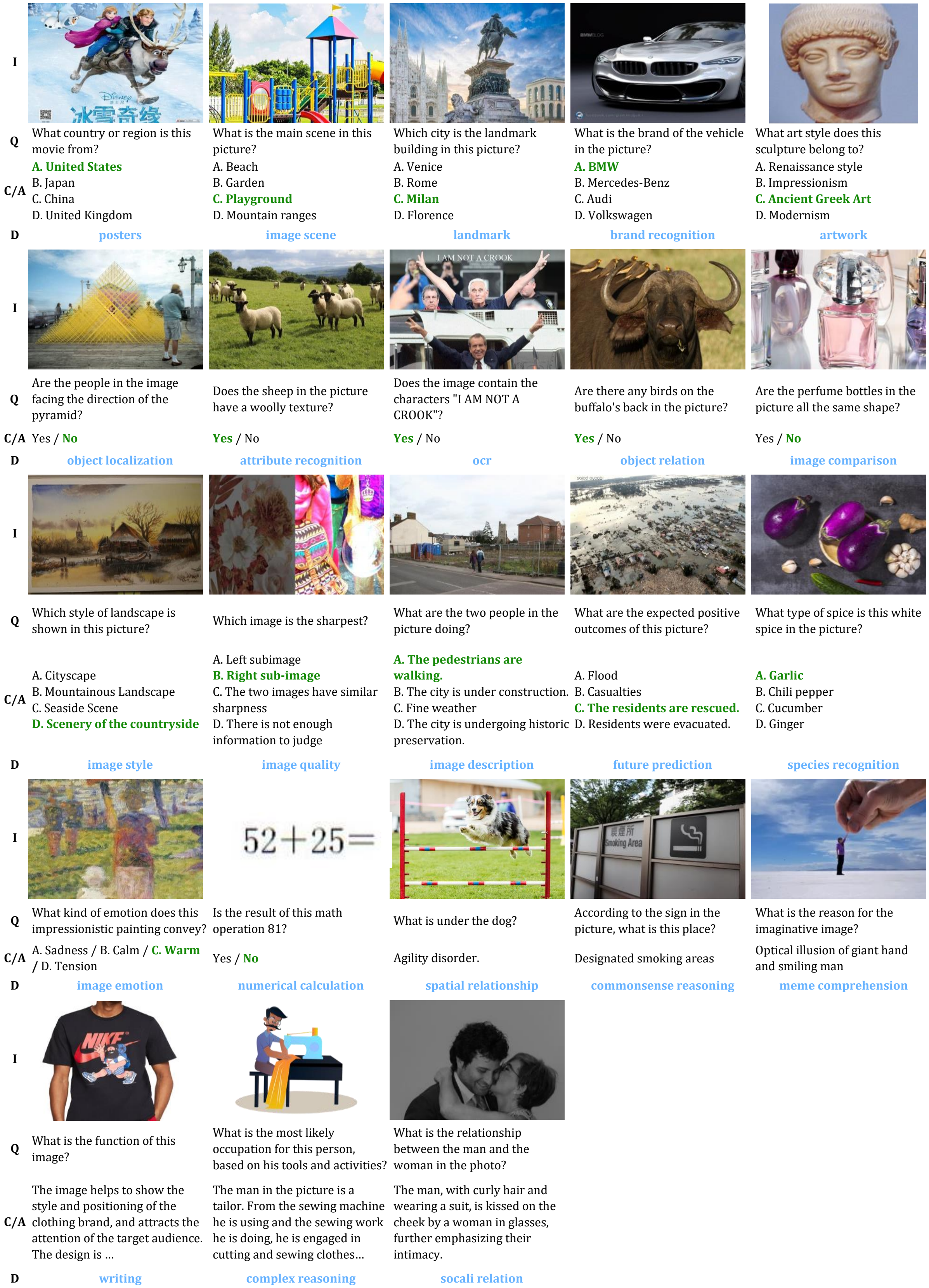


Figure 3 Examples of different domains in MMInstruct. **I** denotes Image. **Q** denotes instruction generated by GPT. **C/A** denotes options and the correct answer to the related instruction; in judgment and multiple-choice questions, the green option indicates the correct answer. **D** denotes the domain.

Table 3 Domain partitioning details of MMInstruct. It includes 23 single turns and 1 multi-round long visual question answering.

Conv Type	Domains
Single-Turn (Perception)	image style, image scene, image quality, image comparison, object localization, object relation, attribute recognition, image description, OCR, posters, artwork, landmark, spatial relationship, brand recognition, species recognition
Single-turn (Reasoning)	numerical calculation, image emotion, commonsense reasoning, complex reasoning, social relation, future prediction, meme comprehension, writing
Multi-Round	multi-round long visual question answering

MultiInstruct [75] introduces the first human-label visual instruction tuning dataset. Mini-GPT4 [82] generated its instruction-based dataset by composing image-text datasets and handwritten instruction templates. LLaVA [41] employs ChatGPT/GPT-4 to convert image-text pairs into multi-modal instruction-following data. Subsequently, several instruction datasets (*e.g.*, LAMM [77], MIMIC-IT [28], and Macaw-LLM [48]) further encompass 3D-domain, Audio and videos examples for instruction tuning. InstructBLIP [15] and LLaVA-1.5 [40] incorporate academic-task-oriented Visual Question Answering (VQA) datasets to augment the model's visual capabilities. M^{3}IT [34] further scaled up the instruction data to 2.4 million instances.

Some works focus on improving the performance of VLLMs in specific domains or emphasize enhancing certain aspects of the model's capabilities. VideoChat [33], TimeChat [56], and Valley [47] build video-centric instruction datasets aimed at enhancing the video comprehension, conversation, and complex reasoning capabilities of VLLMs. ScienceQA [45] and MMMU [80] construct question-answer pairs from primary and secondary school classes and college exams, respectively, covering diverse disciplines and emphasizing perception and reasoning with domain-specific knowledge. LLaVAR [81] augments visual instruction tuning with text-rich images using OCR tools and GPT-4. Some datasets (*e.g.*, mPLUG-DocOwl [76], InstructDoc [61]) focus on document understanding tasks, necessitating models to possess robust OCR capabilities as a foundation. LRV-Instruction [39] includes both positive and negative instructions to mitigate hallucination, resulting in a more robust model. Shikra [9] and All-Seeing [68,69] utilize data with region annotations to enhance the referential dialogue and panoptic visual recognition and understanding capabilities of VLLMs. Vision-Flan [74], consisting of 22 tasks drawn from academic datasets, is built to address issues of poor generalization, hallucination, and catastrophic forgetting in models trained on GPT-4 synthesized data.

Compared to the previous works, we aim to construct a visual instruction tuning dataset that encompasses a wider range of domains, features more precise annotations, and provides richer question-answering forms and types.

3 Method

In this paper, we propose a visual instruction tuning dataset, named MMInstruct, which ensures diverse images, high annotation quality, and diverse instructions. Our dataset is primarily divided into 24 domains, including 23 single-turn question-answering domains and one multi-round long visual question-answering domain. The partitioning details are shown in Table 3. MMInstruct comprises a total of 161K high-quality detailed image captions and 973K instruction data.

To overcome the high cost of dataset construction while increasing dataset coverage and diversity, we propose a semi-automatic and low-cost instruction generation data engine utilizing GPT-4V, GPT-3.5 and manual correction, as shown in Figure 4. Our data engine comprises six steps: (a) Image Collection, (b) Image Caption Generation, (c) Seed Question Collection, (d) Automatic Instruction Generation, (e) Dataset Expansion, and (f) Manual Correction. Initially, we collect a large and diverse set of images from various sources and employ GPT-4V to generate detailed image captions. Seed questions are curated by our experts and validated for effectiveness. Subsequently, leveraging both the image captions and seed questions, GPT-3.5 automatically generates a rich and diverse set of instruction data. Additionally, we employ various methods to expand our dataset. Finally, manual corrections are made to ensure data quality and accuracy.

Our efforts primarily revolve around the following: (1) **Image Diversity**: Since high-quality images are difficult to obtain, image acquisition in existing instruction datasets mostly relies on open-source image datasets, but this also limits the scope of image inclusion. Therefore, we propose a process to

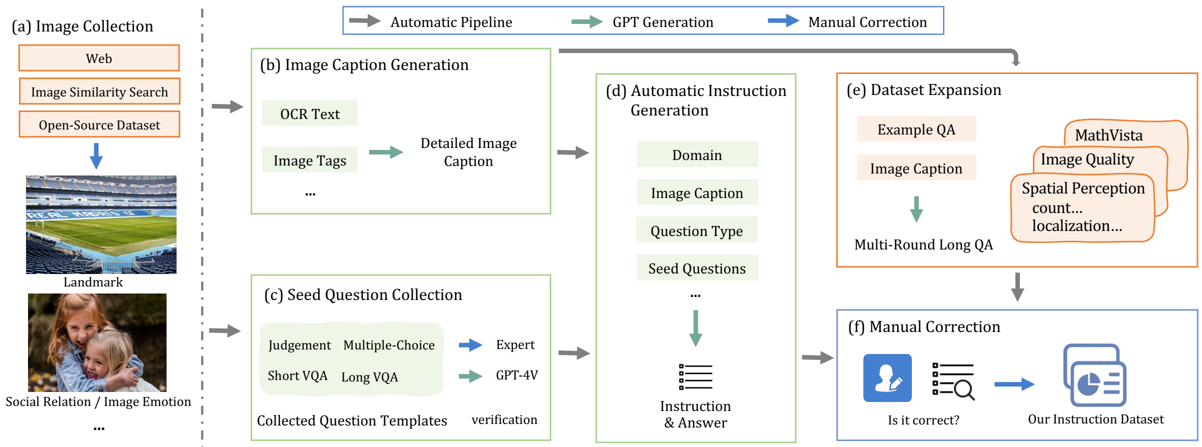


Figure 4 Data engine for MMInstruct. Our data engine consists of automatic generation and manual correction. (a) We collect a large number and diversity of images from a variety of sources. (b) We utilize GPT-4V to generate detailed image descriptions based on the image and context of the image. (c) Human experts collect seed questions and validate the effectiveness of seed questions using GPT-4V. (d) Then, leveraging those detailed image descriptions and seed questions, we employ GPT-3.5 to generate Instruction-Answer pairs. (e) We also use several methods to expand our dataset. (f) Finally, additional manual corrections are performed.

Table 4 Some key phrases used for searching images on the web.

Domains	Search key phrases
Feature Prediction	typhoon, traffic accident, football match, dance, rocket launch, sunrise...
Species Recognition	mammals, marine life, reptiles, insect, virus, plants, fruits...
Meme Comprehension	pepe the frog, confession bear meme, bad luck brian, doge meme...

quickly and extensively collect images from the Internet according to a specific domain, with multiple manual screenings to ensure image quality, as introduced in Section 3.1. (2) **Annotation Quality**: Existing datasets typically rely on existing annotations of images for instruction generation. Rough annotations can cause hallucination problems. Therefore, in Section 3.2, we propose leveraging GPT-4V to obtain rich semantic information from images, followed by manual corrections to ensure annotation quality. (3) **Instruction Diversity**: For a specific domain, users have various instructions. In Section 3.3, we propose compiling a seed question set by analyzing common user instructions. For each image, diverse instructions of four types are generated using the generation pipeline outlined in Section 3.4. Additionally, in Section 3.5, we generate multi-round Long VQA instructions and incorporate other open-source datasets to further supplement our dataset.

3.1 Image Collection

In order to effectively reduce costs while ensuring image diversification, we propose an efficient image collection process. Firstly, experts define key phrases for each domain based on which a large number of open images are crawled from the web, and preliminary screening is conducted on the crawled results. In Table 4, a small portion of the utilized search key phrases is listed. Next, leveraging existing images as a foundation, by utilizing k-Nearest Neighbors image similarity search in the large-scale image repository Laion-5B [57], rigorously assessing their suitability and quality. Note that in order to avoid duplication of images, we strictly deduplicate them through image annotation information and manual screening. Finally, we organize the existing images and select some images from open-source datasets as a final supplement. Through this approach, we collect a total of 161,000 high-quality images across the 24 domains.

3.2 Image Caption Generation

Previous research [39, 41, 81] has relied on annotation data, including scene captions, bounding boxes, and OCR, to depict images and generate instructions for text-only GPT models. However, those poor annotations have become a bottleneck in generating instructions. Therefore, in our data engine, we use

Table 5 Prompt used for detailed image description generation. **Universal** represents the fundamental prompt applicable to all domains. The remaining lines enumerate additional prompt content added for specific domains. *(if has)* means the content is applicable to the corresponding situation.

Domains	Prompt
Universal (For all domains)	<Image> I will provide you an image and related information about the image... Describe the image in as much detail as possible. Image related information: <Image Tag> Text information in the image: <OCR Text> *(if has)* <Special Requirements> *(if has)*
Numerical Calculation	Note that the image provides mathematical problems that may involve numerical values, mathematical formulas, or graphics.
Brand Recognition	Try to identify the brand of the item in the image.
Posters	Try to identify which file/TV show the image comes from.
Landmark	Try to identify the landmark building or place in the image.
Meme Comprehension	Try to discern the intriguing aspects within the image.
Social Relation	Try to identify the relationship between the people in the image.
Spatial Relationship	Try to identify the spatial relationship between the objects in the image.

Table 6 Examples of seed questions in different domains. **General** represents general questions; **Wildcard** represents questions containing placeholders.

Type	Domain	Seed Questions
General	species recognition	Identify the species in the image. What is the scientific name of this species?
	image emotion	Which mood does this image convey? Identify the emotion expressed in this image.
	numerical calculation	Are the calculations in the image correct? Calculate the formulas in the picture.
Wildcard	species recognition	This is an <object>, which species does it belong to? Is the scientific name of this <object> <name>?
	image emotion	Does this image convey the emotion of <specific emotion>? Is the emotion of <some object> in the picture <positive/negative>?
	numerical calculation	What is the <area/volume> of <the geometry> in the image? What should the value of <variable> in the picture be equal to?

GPT-4V to generate a sufficiently detailed and domain-specific image caption for each image. Rich image information can make instruction generation more diverse and reduce hallucination problems.

It is worth noting that we have also implemented the following measures to ensure the accuracy and coherence of the captions generated by GPT-4V. In our caption generation prompt, we additionally integrate domain-specific prior knowledge. For example, in the OCR domain, text recognition is initially performed on images using Google OCR. And we modify the prompt appropriately for different domains. In addition, the annotations of the collected images themselves are also integrated into the prompt for caption generation. The fundamental prompt for all domains and the addition prompt added for specific domains are listed in Table 5.

3.3 Seed Question Collection

Seed questions, serving as a reference for instruction generation in our data engine, directly influence the effectiveness of generated results. These seed questions should be generic, covering the majority of common instructions users may utilize. Furthermore, for different visual domains, corresponding seed questions should also have different focuses to clarify the context of the questions answered. Our seed question templates can be divided into general questions and wildcard questions, Table 6 illustrates some examples of our seed questions. Even for the same domain, the possible types of seed questions are diverse, including different asking methods and questions with or without wildcards.

When constructing MMInstruct, we aggregate a large amount of instruction data from existing open datasets and real users. Subsequently, experts summarize common questions in each domain based on a

Table 7 Key parts of prompts used for instruction generation in different domains. **With Seed** represents the prompt used when generating instruction data based on seed questions; **No Seed** represents the prompt used when there is no generic question template; **Multi-Round** represents the prompt used to generate multi-round long visual question answering.

Type	Domains	Prompts
With Seed	numerical calculation, attribute recognition, landmark, *etc.*	Given a description of the image and a list of questions, you need to design 3 <Question-type> questions and corresponding answers related to the topic of <Domain>... Google OCR content: <OCR Result> Image description: <Image Caption> Question template: <Seed Questions> You must output the generated questions, options, and answers in the following format...
No Seed	complex reasoning, commonsense reasoning	Given a description of the image, you need to ask 3 <Question-type> questions about the image that can be used in the <Domain> task and generate corresponding answers. You must output the generated questions, options, and answers in the following format...
Multi-Round	multi-round long visual question answering	Pretend that you have "seen" an image, based on the description provided below, now you have two tasks: Create 5 Questions using English: <Requests> Answer the Questions using English: <Requests> Example: <Example QA> Image description: <Image Caption>...

statistical analysis of the data to serve as seed questions. Overall, an average of about 10 seed questions are designed per domain. To ensure the effectiveness of the seed questions, a small batch of instructions is generated before the actual instruction generation process. This preliminary step can be used to verify the generated results and make appropriate modifications to the seed questions.

3.4 Automatic Instruction Generation

After obtaining the seed question and image information, we utilize the text-only GPT-3.5 model to generate instruction data. We separately design generation pipelines for four types of questions: judgment, multiple-choice, and Short VQA and Long VQA. In the generation pipeline, for each image, its detailed caption and prior knowledge of the corresponding domain are used as input, and N are randomly selected from the provided seed questions as references (with $N = 3$ in our paper). Then guide GPT to generate instruction data according to the corresponding domain prompt. In order to enable the model to better distinguish the type of instructions, we add indicative utterances corresponding to the question type after each generated question. For example, "Please choose the most appropriate option" will be added to multiple-choice questions. The key part of prompts used is shown in Table 7 line 2.

It is worth noting that in some domains, seed questions may not be universally applicable to all images. For instance, in the numerical calculation domain, the seed questions for formula calculation and variable solving are distinctly different. To enhance the alignment between images and generated instructions, we categorize and match images with the corresponding seed questions. Moreover, the number of seed questions provided for reference is greater than the number of instructions that need to be generated, which can provide fault tolerance space for GPT, thus reducing the occurrence of unreasonable problems and hallucinations.

In commonsense reasoning and complex reasoning domains, there are diverse ways of questioning, hence we haven't collected seed questions. Instead, We employ a prompt for problems directly generated without a universal question template to instruct GPT in directly generating domain-relevant questions from detailed descriptions of the image. The key part of prompts is shown in Table 7 line 3.

Due to factors such as language, culture, and individual habits, user instructions tend to be diverse. Therefore, we encourage the instruction questions generated by GPT to be of various styles, as long as they are semantically close to the seed question. To effectively mitigate hallucination during the generation process, we strictly enforce GPT to generate both questions and their answers at one time. The answers to questions must be correct and explicitly derived from the image information. In particular, for multiple-choice questions, GPT is also required to provide the four options corresponding to the question, ensuring that exactly one option is correct.

3.5 Dataset Expansion

In order to further expand the diversity and versatility of MMInstruct, we also expand the dataset through other methods. On the one hand, we build a similar pipeline to generate multi-round long visual question answering (Multi-Round Long VQA) data to extend the instruction type. On the other hand, we screen and process some data from open-source datasets to extend the domain of our dataset.

Multi-Round Long VQA Data Expansion. In practical user usage, multiple rounds of contextually related questions and answers are a common interaction mode. Multi-Round Long VQA data with rigorous logic and reasonable inference is crucial for model training. Such data aids in the learning of deeper semantic comprehension and inference capabilities, enabling models to perform more accurately and naturally in understanding questions and deducing answers. Therefore, we propose an automated pipeline for constructing Multi-Round Long VQA instructions, leveraging the powerful reasoning capabilities of GPT-4V. Similar to the instruction construction pipeline outlined in Section 3.4, it also utilizes detailed image captions and prior knowledge as input. For the pipeline prompt, while strictly constraining GPT to adhere to the given information, we request it to generate 5 questions along with their corresponding correct answers each time. And these questions should have a continuous logical linkage and evolution between them. The key part of prompts used for Multi-Round Long VQA is shown in Table 7 line 4. It is worth noting that our multi-round response data is longer and more informative than other datasets. This will force the model to have a deeper understanding and rigorous analysis of the questions.

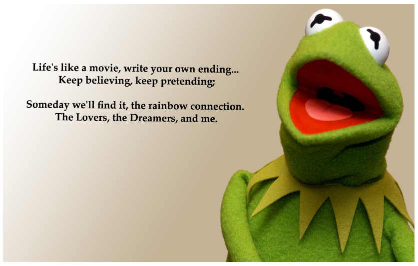
Other Source Data Expansion. In order to further enrich the domain categories contained in our dataset and increase the diversity of instruction formats, we select some data from open-source datasets. This includes mathematics datasets [5, 7, 36, 38, 44, 46, 58], charts and plots [24, 25], scientific figure [26] and map chart [6]. We then convert them into dialogue format and add them to MMInstruct.

3.6 Manual Correction

The instruction data generated by our data engine has basically met the requirements of the instruction dataset. However, in bulk generation, some data inevitably contains problems such as hallucination, grammatical errors, or mismatches between instructions and domain. Therefore, additional manual corrections are necessary for the constructed data, with the cost significantly lower than that of manually constructing a complete dataset. Therefore, it is necessary to perform additional manual corrections on the constructed data, with costs much lower than manually building the dataset from scratch.

In this stage, we provide all data in the form of <*image*, *caption*, *instruction answer*> pair, as shown in Figure 5, for manual correction by multiple professional annotation teams. In order to ensure the quality of the dataset, we set acceptance criteria and hire annotation teams based on the characteristics of the instruction domain. For example, in the OCR domain, in addition to the regular annotation requirements, we additionally require the teams to pay attention to whether the text in the image is correctly recognized, whether the text content is comprehensively recognized, and whether the order of the text output corresponds to the position in the image, and so forth. The teams inspect and modify each instruction data based on the image and caption to meet the strict acceptance criteria. In this process, the same batch of data is shuffled and undergoes three or more rounds of rework. Additionally, humans possess a greater understanding compared to GPT, resulting in manually corrected data exhibiting greater diversity and better alignment with human questioning styles.

Within our data engine, the cost of processing and generating a detailed caption (averaging 200 words) for a 512px×512px image using *gpt-4-1106-vision-preview* is \$0.00885. On this basis, the average cost to generate an instruction using *gpt-3.5-turbo-1106* is \$0.0004, with manual correction costing \$0.13. Therefore, for MMInstruct, handling 161K images and 973K instruction data requires approximately \$128, 304. In contrast, if we manually construct the dataset, each instruction costs around \$0.84, resulting in a total cost of approximately \$817, 320 for MMInstruct. Remarkably, leveraging our data engine costs only one-sixth of what it would cost to build it entirely by hand. This effectively demonstrates the cost-effectiveness of our data engine.



Contexts

Bing Tag: Kermit the Frog

Google OCR: Life's like a movie, write your own ending...keep believing, keep pretending; Someday we'll find it, the rainbow connection. The lovers, the Dreamers and me.

Response (detailed description)

This image features a central animated character —— a frog, likely the character Kermit from the famous "The Muppet Show". This character appears very contemplative and sensitive, fitting well with the textual information in the background. The text is full of sentiment and courage, quite poetic and inspirational, encouraging people to live their life stories as if they were creating a movie script, continuously believing and pretending until they find the legendary "rainbow connection", which is a metaphor for hope or dreams. The mention of "lovers, dreamers, and me" at the end conveys a sense of inclusiveness and resonance, implying that everyone has the right and ability to pursue dreams and seek beauty in life. Overall, the image and text blend seamlessly, with the presence of Kermit adding a warm and friendly atmosphere, while the broad, stylish words provoke deeper reflections on life. This image is likely to trigger viewers' reflections on their personal dreams and contemplation on how to maintain optimism and creativity in life.

Response (a judgement question answering about OCR)

Question: Does the text on the image "Life's like a movie, write your own ending...Keep believing, keep pretending; Someday we'll find it, the rainbow connection. The lovers, the Dreamers, and me."?
Answer: Yes

Response (a multiple-choice question asnwering about meme comprehension)

Question: What does this image usually represent?
A. Humor and fun
B. Sorrow and melancholy
C. Courage and Change
D. Inclusiveness and resonance
Answer: D. Inclusiveness and resonance

Figure 5 One example of the generated instruction data using our data engine. The top block shows the contexts such as the image tag from Bing and the OCR results obtained from Google, those contexts are used to prompt GPT to generate a detailed description of the given image. The second block is a detailed image description generated based on the image and context. The last two blocks show two different types of instruction data generated based on detailed image descriptions and seed questions.

4 Experiments

4.1 Experiment Setup

To verify the effectiveness of our proposed dataset MMInstruct, we conduct a series of evaluation experiments. We follow the design of the advanced VLM architecture LLaVA-1.5 [40], which mainly consists of three parts: the pre-trained vision encoder CLIP-ViT-L-336px [54], the pre-trained large language model Vicuna-v1.5 [14], and a 2 layer MLP projection.

Training Details. We adopt the same two-stage training design as LLaVA-1.5. During the pre-training stage, we keep the vision encoder and large language model frozen and use the LCS-558K pre-training dataset to train the MLP projection. In the fine-tuning stage, we keep the vision encoder frozen and combine the LLaVA-665K instruction dataset with our MMInstruct to fine-tune the MLP projection and large language model. Meanwhile, we use the same hyper-parameters as LLaVA-1.5.

Evaluation Benchmarks. We evaluate our visual instruction tuning dataset MMInstruct using the same benchmarks as LLaVA-1.5, including traditional academic visual question answering benchmarks: **VQAv2**: VQA^{v2} [19], **GQA** [21], **VizWiz** [20], **ScienceQA-Img**: SQA^{I} [45], and **TextVQA**: VQA^{T} [59]; comprehensive multi-modal evaluation benchmarks: **POPE** [35], **MME** [18], **MMbench**: MMB [42], **MMbench-Chinese**: MMB^{CN}, **SEED-Bench**: SEED [29], **LLaVA-Bench** (In-the-Wild): LLaVA^{W} [41] and **MM-Vet** [79].

Table 8 Comparison with state-of-the-art VLLMs on traditional VQA benchmarks. Priv: the data are private. * denotes the training images of the datasets are observed during training. The best results are marked in **bold**, and the second best results are underlined.

Method	LLM	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}
InstructBLIP [15]	Vicuna-7B	–	49.2	34.5	60.5	50.1
IDEFICS-9B [22]	LLaMA-7B	50.9	38.4	35.5	–	25.9
Qwen-VL [2]	Qwen-7B	78.8*	59.3*	35.2	67.1	**63.8**
Qwen-VL-chat [2]	Qwen-7B	78.2*	57.5*	38.9	68.2	<u>61.5</u>
LLaVA-1.5 [40]	Vicuna-7B	78.5*	62.0*	50.0	66.8	58.2
LLaVA-1.5 +MMInstruct (ours)	Vicuna-7B	<u>79.6</u>*	<u>62.8</u>*	53.3	71.2	57.7
BLIP-2 [31]	Vicuna-13B	65.0	41.0	19.6	61.0	42.5
InstructBLIP [15]	Vicuna-13B	–	49.5	33.4	63.1	50.7
IDEFICS-80B [22]	LLaMA-65B	60.0	45.2	36.0	–	30.9
Shikra [9]	Vicuna-13B	77.4*	–	–	–	–
LLaVA-1.5 [40]	Vicuna-13B	**80.0***	**63.3***	<u>53.6</u>	<u>71.6</u>	61.3
LLaVA-1.5 +MMInstruct (ours)	Vicuna-13B	**80.0***	62.6*	**55.8**	**74.2**	60.9

Table 9 Comparison with state-of-the-art VLLMs on recent Multi-modal benchmarks.

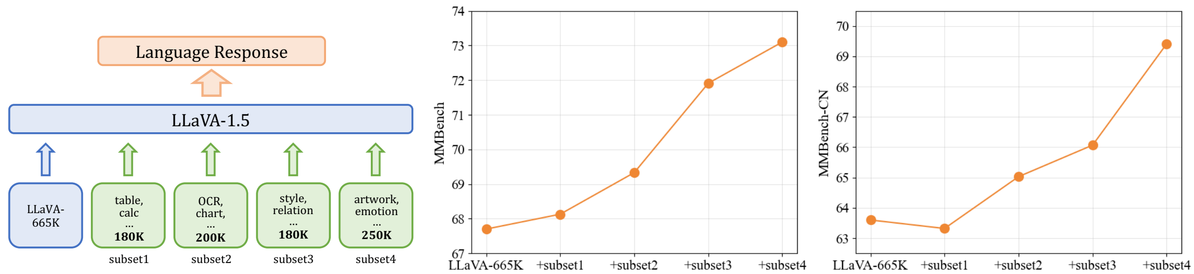
Method	LLM	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MM-Vet
InstructBLIP [15]	Vicuna-7B	–	–	36.0	23.7	53.4	60.9	26.2
IDEFICS-9B [22]	LLaMA-7B	–	–	48.2	25.2	–	–	–
Qwen-VL [2]	Qwen-7B	–	–	38.2	7.4	56.3	–	–
Qwen-VL-chat [2]	Qwen-7B	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5 [40]	Vicuna-7B	85.9	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 +MMInstruct (ours)	Vicuna-7B	<u>86.8</u>	<u>1591.8</u>	<u>69.8</u>	<u>65.2</u>	<u>64.6</u>	<u>71.0</u>	34.4
BLIP-2 [31]	Vicuna-13B	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP [15]	Vicuna-13B	78.9	1212.8	–	–	–	58.2	25.6
IDEFICS-80B [22]	LLaMA-65B	–	–	54.5	38.1	–	–	–
Shikra [9]	Vicuna-13B	–	–	58.8	–	–	–	–
LLaVA-1.5 [40]	Vicuna-13B	85.9	1531.3	67.7	63.6	61.6	70.7	<u>35.4</u>
LLaVA-1.5 +MMInstruct (ours)	Vicuna-13B	**86.9**	**1626.2**	**72.1**	**68.0**	**64.7**	**74.5**	**37.9**

4.2 Main Results

As shown in Table 8 and Table 9, in quantitative comparisons with leading VLLMs, our 7B and 13B models significantly outperform LLaVA-1.5 models across various benchmarks, including both academic visual question answering and multi-modal evaluation benchmarks. It is worth noting that our 13B model achieves state-of-the-art performance on 10 out of 12 benchmarks.

Results of Visual Question Answering Benchmarks. On general VQA benchmarks, Our 13B model has shown significant improvements over LLaVA-1.5 models, particularly in VizWiz and ScienceQA, especially ScienceQA exhibiting an improvement of nearly 3% compared to LLaVA-1.5. Additionally, our model has demonstrated competitive performance in VQAv2, GQA, and TextVQA benchmarks as well.

Results of Multi-modal Benchmarks. In recent comprehensive multi-modal benchmarks, which contain fine-grained multi-modal tasks across a wide range of tasks. Our model achieves state-of-the-art performance on these benchmarks, surpassing LLaVA-1.5 comprehensively. Specifically, we achieve a substantial gain of 94.9 points (1626.2 vs. 1531.3) on MME and an impressive improvement of 4.4 points (68.0 vs. 63.6) on MMBench-CN. Furthermore, our model exhibits significant enhancements over LLaVA-



(a) Dataset partitioning by domain (b) Impact on MMBench (c) Impact on MMBench-CN

Figure 6 Performance on MMBench/MMbench-CN versus the number of training domains.

1.5 on other multi-modal benchmarks, including POPE, MMBench, SEED, LLaVA-Bench (In-the-Wild), and MM-Vet. These results highlight the effectiveness of MMInstruct.

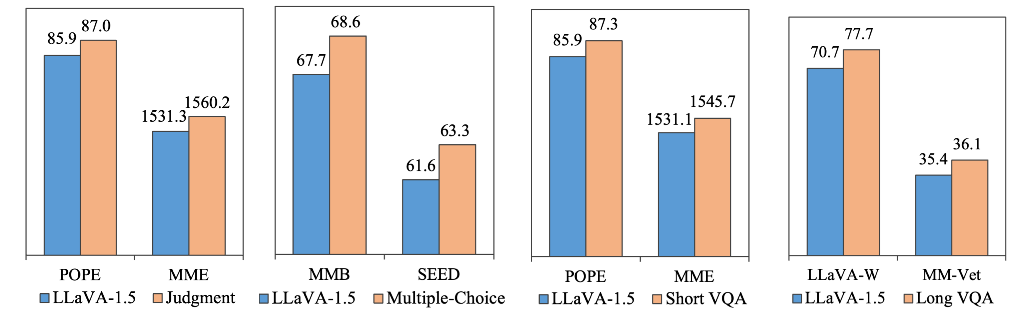
We attribute these performance improvements to the advantages of MMInstruct, including image diversity, instruction diversity, and high-quality annotations. The dataset includes images from a broader range of domains, enabling the model to have greater generalization capability. Simultaneously, detailed and precise image captions and strictly set prompts when generating instructions effectively reduce model illusion and deviation questions. Moreover, our data engine ensures that the generated instructions encompass the four common question types and feature complex and varied sentence structures. This compels the model to deeply understand the essence of tasks rather than merely learning surface-level sentence patterns. Consequently, it exhibits satisfactory responses to different instruction formats across VQA and multi-modal benchmarks.

4.3 Ablation Studies

While keeping pre-training data the same, we also report the quantitative results of varying the fine-tuning data. This can provide insights for more efficient use of MMInstruct later.

Effect of Domain Diversity in MMInstruct. To investigate the impact of domain diversity on model performance, we conduct ablation experiments on the number of domains. We randomly divide the data of MMInstruct into subsets with similar data size by domain. Subsequently, we employ the LLaVA-1.5 13B model pre-trained with the same settings. Only during fine-tuning stages do we incrementally add new data subsets for each experiment. Figure 6 shows the relationship between the performance of the two comprehensive evaluation benchmarks, MMBench and MMBench-CN, and the number of MMInstruct fields used in the instruction fine-tuning stage. It's evident that increasing the number of domains can significantly improve the performance of the model in both Chinese and English. Furthermore, this performance improvement shows a linear increasing trend overall. These findings robustly confirm the rationality of our domain categorization and validate the effectiveness of domain diversity.

Effect of Question Types Diversity in MMInstruct. We analyze the impact of different question types used in the instruction fine-tuning stage on model performance. The model used in the experiment is LLaVA-1.5 13B, which is pre-trained without any instruction fine-tuning. We categorize the data in MMInstruct based on question types into judgment, multiple-choice, Long VQA, and Short VQA. In each experiment, we randomly select 150K samples from only one type of question data and combine them with the LLaVA-665K instruction dataset to fine-tune the model. In comparison to the baseline utilizing solely the LLaVA-665K dataset, the four sub-figures depicted in Figure 7 showcase two prominent benchmarks for distinct types of questions each. Specifically, the inclusion of Long VQA data leads to substantial improvements of 7.0 and 0.7 points in the LLaVA-Bench (In-the-Wild) and MM-Vet benchmarks, respectively. The utilization of multi-choice data significantly enhances the MMBench and SEED benchmarks by 0.9 and 1.7 points, respectively. This strongly indicates that the diversity of question types is crucial for models to effectively comprehend tasks.



(a) Impace after adding Judgement questions (b) Impace after adding Multiple-Choice questions (c) Impace after adding Short VQA Questions (d) Impace after adding Long VQA Questions

Figure 7 Performance on multi-modal benchmarks versus the question types.

Effect of Multi-Round Long VQA in MMInstruct. As shown in Table 10, we attempt to remove the Multi-Round Long VQA data from MMInstruct to assess its impact on the model. For a fair comparison, we train the model using the same settings and images, with the only difference being that the Multi-Round Long VQA instructions are removed from MMInstruct for all images. The results show that Multi-Round Long VQA data can make the model achieve significant gains, with improvements of 2.2 and 0.5 points on the LLaVA-Bench (In-the-Wild) and MM-Vet benchmarks, respectively. We speculate that this is because the Multi-Round Long VQA data effectively trains the model to handle long contexts, while the generation of long outputs also effectively enhances the model's inference capabilities for complex tasks. This proves the necessity of adding high-quality Multi-Round Long VQA data to the instruction fine-tuning dataset.

Table 10 Performance on LLaVA-Wild and MM-Vet versus Multi-Round Long VQA. **NoMR** represents a model tuned without Multi-Round Long VQA.

Method	LLaVA^{W}	MM-Vet
LLaVA-1.5	70.7	35.4
LLaVA-1.5 +MMInstruct-NoMR (ours)	72.3	37.4
LLaVA-1.5 +MMInstruct (ours)	**74.5**	**37.9**

4.4 Visualizations

We visualize some example outputs of LLaVA-1.5 and our model in Figure 8. Note that the text marked with the corresponding background color represents the correct reasoning process and results, while the text marked in red indicates hallucinations. We can see that with additional instruction fine-tuning on MMInstruct, our model is better able to reason with contextual information in Multi-Round Long VQA, and its answers contain a more explicit reasoning process. In particular, our model is able to more accurately recognize text in pictures and understand the spatial relationships of image objects, which is significantly beneficial for logical reasoning.

5 Conclusion

In this paper, we constructed a high-quality, diverse visual instruction tuning dataset MMInstruct, which consists of 973K instructions from 24 domains. Specifically, MMInstruct contains diverse question forms and types, including Judgment, Multiple-Choice, Long VQA, and Short VQA. To construct MMInstruct, we propose an instruction generation data engine that leverages GPT-4V, GPT-3.5, and manual correction. Our data engine enables semi-automatic, low-cost, and multi-domain instruction generation. Compared to manual construction, our data engine's cost is only 1/6 of manual annotation while ensuring annotation quality and data diversity. Then, we incorporate MMInstruct into the instruction fine-tuning phase of LLaVA-1.5 to evaluate its effectiveness. The results show that our model demonstrates impressive performance across multiple multi-modal benchmarks. Additionally, we also perform comprehensive ablation experiments to analysis the impacts of varying the fine-tuning data on VLLMs. These results clearly demonstrate that MMInstruct benefits visual instruction tuning.

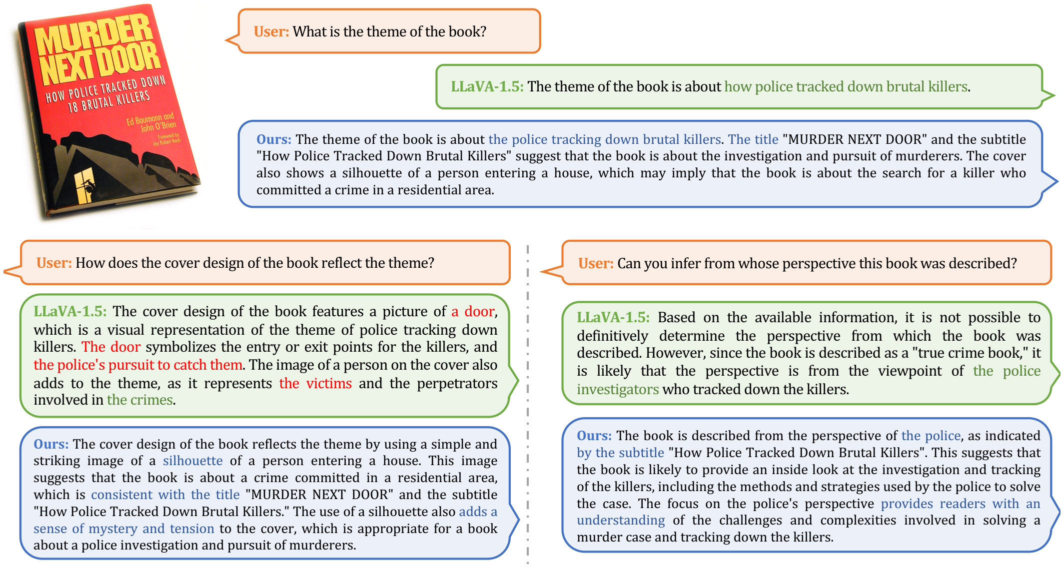


Figure 8 Visualization of outputs comparison between LLaVA-1.5 and our model.

References

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

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

3 Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022.

4 Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining, 2022.

5 Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, 2022.

6 Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

7 Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

8 Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022.

9 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.

10 Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

11 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. In *ICLR*, 2022.

12 Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

13 Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

14 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)*, 2(3):6, 2023.

15 Wenliang Dai, Junnan Li, Dongxu Li, AnthonyMeng Huat, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

16 Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.

17 Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

18 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.

19 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

20 Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

21 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

22 IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023.

23 Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, 2021.

24 Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

25 Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

26 Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.

27 Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

28 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*, 2023.

29 Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

30 Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023.

31 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*, 2023.

32 Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

33 Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

34 Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M ^{3} it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

35 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *EMNLP*, 2023.

36 Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023.

37 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 *ECCV*, 2014.

38 Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

39 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*, 2023.

40 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

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

42 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*, 2023.

43 Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

44 Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

45 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. *NeurIPS*, 2022.

46 Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.

47 Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

48 Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

49 OpenAI. Chatgpt, 2022.

50 OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

52 TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022.

53 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.

54 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

55 Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

56 Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv*, abs/2312.02051, 2023.

57 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. *NeurIPS*, 2022.

58 Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015.

59 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

60 Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

61 Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *AAAI*, 2024.

62 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

63 InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM, 2023.

64 Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

65 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.

66 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

67 Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

68 Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024.

69 Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023.

70 Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, 2023.

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

72 Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2, 2022.

73 Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024.

74 Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*, 2024.

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

76 Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding, 2023.

77 Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

78 Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

79 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

80 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

81 Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

82 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.

83 Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe:

Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*, 2022.
84 Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv preprint arXiv:2112.01522*, 2021.