

Oasis: One Image is All You Need for Multimodal Instruction Data Synthesis

Letian Zhang^{*1,2} Quan Cui² Bingchen Zhao³ Cheng Yang²

¹Tongji University ²Bytedance ³University of Edinburgh

Abstract

The success of multi-modal large language models (MLLMs) has been largely attributed to the large-scale training data. However, the training data of many MLLMs is unavailable due to privacy concerns. The expensive and labor-intensive process of collecting multi-modal data further exacerbates the problem. Is it possible to synthesize multi-modal training data automatically without compromising diversity and quality? In this paper, we propose a new method, **Oasis**, to synthesize high-quality multi-modal data with only images. **Oasis** breaks through traditional methods by prompting only images to the MLLMs, thus extending the data diversity by a large margin. Our method features a delicate quality control method which ensures the data quality. We collected over 500k data and conducted incremental experiments on LLaVA-NeXT. Extensive experiments demonstrate that our method can significantly improve the performance of MLLMs. The image-based synthesis also allows us to focus on the specific-domain ability of MLLMs. Code and dataset are publicly available at https://github.com/Letian2003/MM_INF.

1. Introduction

Multi-modal large language models (MLLMs) have become a popular research topic in recent communities due to their superior performance in various multi-modal tasks. The success of MLLMs relies heavily on the large-scale training data, which directly compose the model’s knowledge base. However, the lack of multi-modal training data has been a bottleneck for the development of MLLMs, since the training data of top MLLMs are typically private. Therefore, an effective way to synthesize high-quality multi-modal data has been a long-standing challenge for the community.

Previous studies have presented some effective methods to synthesize multi-modal data with low cost. LLaVA [24] takes a GPT-assisted method to generate multi-modal instruction-following data based on existing image-pair data. ALLAVA [4] achieves data synthesis using a captioning-then-QA fashion with the assistance of GPT-4V.

*Work done during internship at Bytedance.

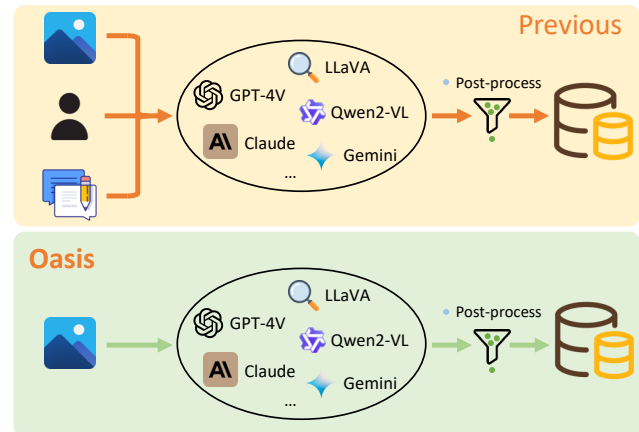


Figure 1. **Comparison of previous methods and our proposed Oasis framework for multi-modal data synthesis.** Previous approaches rely on an input image, complex prompts for generating text, and human labor. Responses are generated from advanced LLMs and MLLMs (e.g., GPT-4V [1] and Qwen2-VL [2, 40]). Interestingly, our proposed **Oasis** requires only a single image to generate multi-modal instruction-following data, showing great simplicity and practical value.

Some recent research focuses on the diversity and complexity of data. For example, MMEvol [29] is an image-text instruction evolution framework, which iteratively enhances instruction diversity in multiple designed domains.

We analyze the existing synthesis approach and concluded three main concerns across this line of work: **(1) Constant pipeline compromises data diversity.** Invariable prompts and fixed flow constrains the data scope and homogenizes the difficulty, leading to an incomprehensive model. **(2) Insufficient control on data quality.** It is hard to synthesize high-quality data which could significantly improve the representation ability of MLLMs. Most studies adopt a caption-based GPT generation strategy to mitigate this problem. **(3) Complicated framework involves human engagement.** A self-contained synthesis framework typically requires human efforts to design data patterns or prompts, which makes data synthesis troublesome.

In response to these concerns, we propose **Oasis (One image is all you need for multimodal data synthesis)**, a novel and straightforward method to create high-quality and

diverse multi-modal data with only images. Inspired by Magpie [43], we break the traditional input tokens and entice a strong MLLM to generate self-aligned instructions. No single text prompt is required and the only input to the MLLM is the image, which could be easily obtained from the web. The auto-regressive nature of MLLMs leads them to generate diverse instructions based on their own knowledge base. We dive into the property of good instructions and carefully design several standards to filter out low-quality data.

Oasis is a simple yet effective method for multi-modal data synthesis that takes only visual content as prior knowledge. This inherent image-based nature makes the generated data domain heavily dependent on the image domain. We take advantage of this characteristic and produce domain-oriented multi-modal data by controlling the source of images, while not compromising the data quality and diversity. To validate the effectiveness of our method, we collect over 500k **Oasis** data and conduct extensive experiments on LLaVA-NeXT [23]. The results demonstrate that incorporating our synthesized data into the original training set significantly enhances MLLM performance across 14 benchmarks with different backbones. Moreover, our method outperforms existing synthesis approaches by a large margin. We conduct a series of ablation studies to further verify the effectiveness of our quality control strategy. Additionally, we perform a case study in the OCR domain to showcase our method’s capability in domain-specific tasks.

We summarize our contributions as follows:

1. A novel and straightforward method **Oasis** is proposed to synthesize multi-modal data of high diversity and quality, which only requires images as input.
2. By concluding the property of high-quality multi-modal data, we handcraft an array of quality control techniques to ensure the data quality.
3. Extensive experiments demonstrate that **Oasis** data effectively enhance MLLM capabilities across different backbones and outperform other synthesis methods.
4. Over 500k **Oasis** data and consequent models will be made publicly available, hoping to facilitate future research in this field.

2. Related Work

2.1. Multi-modal Large Language Models

Multi-modal large language models (MLLMs) have achieved remarkable success in various vision-language tasks, *e.g.*, GPT-4o [1], LLaVA series [19, 22–24], InternVL series [7–9] and Qwen-VL series [2, 38, 40, 44]. Popular MLLM architecture is composed of three components: a vision encoder, a large language model, and a projector. Contrastive Language-Image Pre-training (CLIP) [34] vision encoders are widely used and incorporated with diverse

LLMs [1, 11, 13, 38, 44]. The projector can be a compact module based on cross-attention layers or MLPs. The multi-modal pre-training paradigm has been widely studied in recent years. A common multi-modal training process [39] is typically divided into two stages: pre-training and fine-tuning. In the pre-training phase, the model is trained on a large-scale multi-modal caption dataset, and the objective is to align representations of different modalities. The pre-training dataset is usually collected from the web and mainly consists of captions. In the fine-tuning stage, the model is fine-tuned on a specific downstream task, which involves a large amount of instruction-following data. The multi-modal pre-training paradigm has been proven to be effective in improving MLLM performance on various vision-language tasks, *e.g.*, general question answering, OCR and visual reasoning.

2.2. Multi-modal Data Synthesis

The lack of multi-modal training data has been a bottleneck for the development of MLLMs. To address this issue, researchers have proposed various methods to synthesize multi-modal data in both LLM and MLLM communities [5, 10, 14, 21, 26, 37, 41, 43, 48].

Multi-modal synthesis methods can be roughly divided into two categories: rule-based methods and generation-based methods. Rule-based methods synthesize multi-modal data based on predefined rules, while generation-based methods generate multi-modal data using generative models. Rule-based methods are usually simple and efficient, but they may lack diversity. Data synthesis methods in the multi-modal domain often take inspiration from the works focusing on generating language-only instruction data, example works include [10, 37, 41, 42], leveraging the capability of the GPT4 model, Alpaca and Vicuna models [10, 37] leverage the self instruct method to generate a large number of instruction data to improve the instruction following ability of base models. Evol-Instruct [42] takes a step further by using evolution-based prompts to improve the diversity of the generated instruction data. ALLaVA [4] provides high-quality multi-modal data by handcraft efforts to rewrite and polish instruction-following data. Cambrian-1 [39] revises classical computer vision datasets by designing specific text as instructions. Generation-based methods can generate diverse multi-modal data, but they may require a large amount of computational resources. For example, LLaVA’s authors create LLaVA-Instruct [24] by using advanced LLMs to label data. MMInstruct [26] employs state-of-the-art MLLMs to generate data. MMEvol [29] and VILA2 [12] study the prompt evolution technique to create complex and diverse multi-modal training data. The majority of existing methods involve complex prompt designs and human labor in the synthesis procedure. The most related work to ours is a prompt-free text data syn-

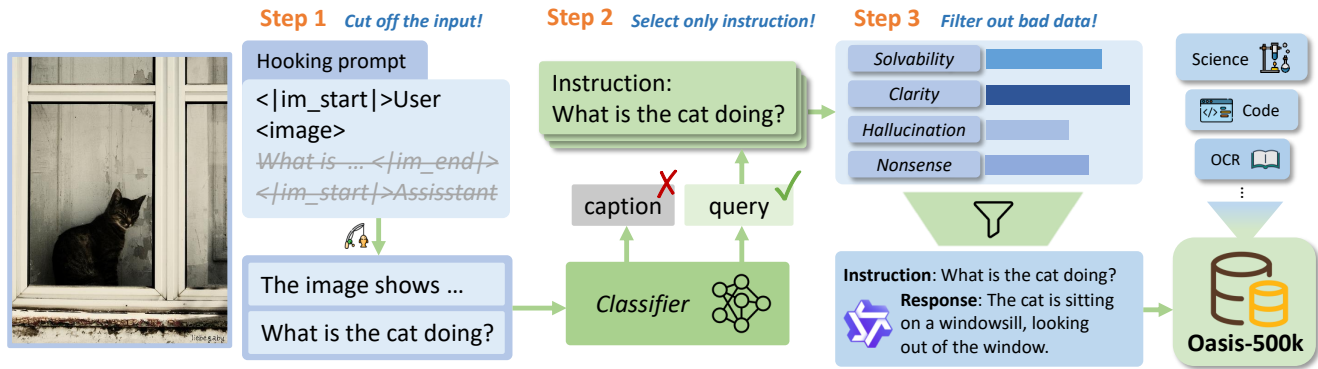


Figure 2. **Detailed Oasis pipeline.** This figure illustrates the full process of data synthesis with **Oasis**. The pipeline consists of three steps: data synthesis, data categorization, and instruction quality control. In Step 1, we break the traditional input tokens and entice a strong MLLM to generate instructions based on the image. In Step 2, we filter out non-instruction-following data by an LLM. In Step 3, a quality control mechanism is proposed to ensure the remained instructions are reasonable and high-quality. **Oasis** exhibits a straightforward and efficient way to synthesize multi-modal training data with low demands (*i.e.*, a single image input). The empirical results show that our method can significantly improve the performance of MLLMs.

thesis work Magpie [43], and further discussion of relations between Magpie and our method is provided in Appendix A.1. Inspired by this study, we propose a simple yet effective method named **Oasis**, which generates multi-modal instruction-following data with a single image.

3. Methodology

3.1. Motivation

Automated multi-modal data generation often suffers from inadequate data quality and diversity, while demanding heavy human involvement. Previous studies have presented impressive methods to eliminate these obstacles. However, we believe that the synthesis process can be performed in a more efficient but straightforward way, without compromising data quality and diversity. In the following, we present a novel method named **Oasis**, which synthesizes multi-modal instruction-following data with only one image.

3.2. Oasis: A Novel Synthesis Method

Our **Oasis** data synthesis pipeline consists of three steps, and the detailed flow is presented in Fig. 2. In the first step, we extract the response of a fully optimized MLLM with a “hooking” prompt. In the second step, we categorize the response of MLLM with the assistance of an LLM. In the third step, we design a quality control mechanism to filter inferior responses with an LLM. Finally, the corresponding answers are generated by an MLLM. In the following, we present the details of the above steps.

Step 1: Data synthesis with “hooking” prompt. We use an MLLM to extract the response to the image input. The typical input to an MLLM can be broken into four components: the pre-query template, the visual content, the instruction, and the post-query template. In the case

of Qwen2-VL, an input can be “`<|im_start|>User <image>Describe the image.<|im_end|> <|im_start|>Assistant`”. Among these components, the pre-query template and the post-query template are fixed, dividing the boundary of the user and the assistant part. Therefore, the response of the MLLM can be formulated as $Resp = \Theta(vision, instruction)$, where Θ represents the MLLM. In **Oasis**, the blue part is considered **a hook** for the red part. As shown in Fig. 2, we intentionally remove the query part and the post-query template, making the MLLM generate the instruction autoregressively based only on the hooking prompt, *i.e.*, $Inst = \Theta(vision)$. By this means, the generated instructions are free from manually designed prompts, thus more diverse.

Step 2: Data categorization. During synthesis, we observe that the generated data can be roughly divided into two categories: instruction-following and caption. The data can be either instruction-following, where the instruction guides the user to perform a specific task, or caption, where the instruction describes the content of the image. This phenomenon can be explained by the interleaved MLLM image-text training process. To select only instruction data, we design a categorization mechanism to classify the data into two categories as depicted in Fig. 2. Specifically, we prompt an LLM as a classifier to predict the category of the data. If the data contains instructions, it is classified as instruction-following data and an instruction is extracted. Otherwise, it is classified as caption data and deserted. We use few-shot to improve the classification accuracy and the full prompts can be found in Appendix B.1.

Step 3: Instruction quality control. With in-depth observation of previous works, we believe that the quality of

Table 1. **Data length statics.** This table presents a comparison of data length statistics between LLaVA-NeXT and **Oasis-500k**. Both instructions and responses in our data are longer and more variable than those in LLaVA-NeXT. This indicates that our data is more informative and diverse.

Data source	Ave. Length		Std. Deviation	
	Inst.	Resp.	Inst.	Resp.
LLaVA-NeXT	45.24	34.16	55.03	185.30
Oasis	76.80	71.16	375.76	529.34

the instruction is the key to the success of the synthesized data. Therefore, after a thorough analysis of the properties of high-quality instructions, we summarize four important characteristics. **(1) Solvability.** Does the image provide all the necessary information to answer the question comprehensively. **(2) Clarity.** How precisely the question conveys its intent and whether it allows for a definitive answer. **(3) Hallucination.** Alignment between the question’s content and the actual content of the image. **(4) Nonsense.** Whether the question is grammatically correct, coherent, and semantically meaningful. We rate all instructions according to these four dimensions on a scale of 1 to 5 and filter out low-quality instructions. The first 3 standards are evaluated by an MLLM, while the last one is evaluated by an LLM, since LLM is more sensitive to the language quality. Specifically, we handcrafted elaborate scoring criteria for each dimension, and the prompts are listed in Appendix B.2. Evaluation cases are provided in Fig. 6c. Response quality control is proven instead to be ineffective in Sec. 4.4.

Oasis-500k data. We select Cambrian-10M [39] dataset as the image source for reproduction, and collect 500k data, named **Oasis-500k**. Due to the great scalability, **Oasis** data could be easily scaled as long as there are enough images. Therefore, the size of the data can grow linearly over time.

3.3. Oasis-500k Data Attributes

In this part, we provide in-depth explorations of the attributes of synthetic data for better understanding **Oasis**.

Oasis data has long text lengths of instructions and responses. In Tab. 1, we report the average length and standard deviation of the instruction and response data in LLaVA-NeXT and **Oasis-500k**. The results show that the average length of **Oasis** data is about double that of LLaVA-NeXT, indicating that our data is more informative and detailed. More details of the input image or more complicated reasoning contents can be provided by **Oasis** data. The standard deviation of **Oasis** data is also higher than that of LLaVA-NeXT, suggesting that our data is more diverse and takes variable forms.

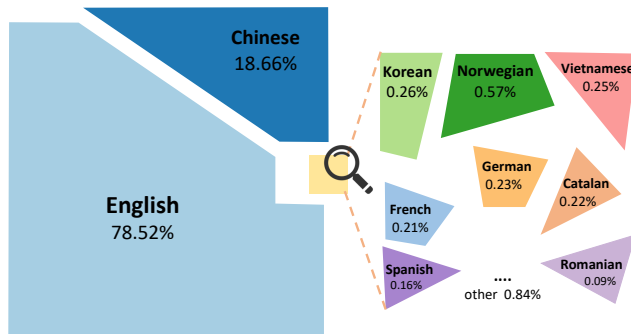


Figure 3. **Language type breakdown.** The distribution of language types in **Oasis** data. English takes up the majority, while other languages are also well-represented. In total, 46 language types are included in the dataset.

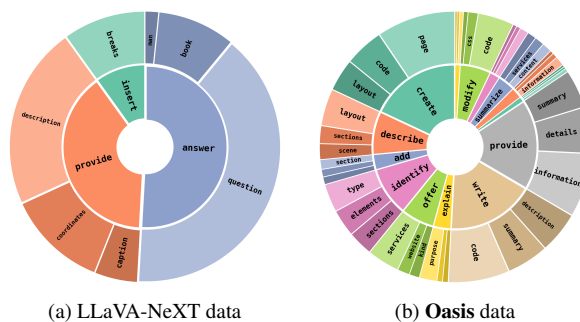


Figure 4. **Root verbs and top noun objects.** The charts show the most common root verbs and their top 3 noun objects in LLaVA-NeXT and **Oasis** data. Word combinations in LLaVA-NeXT data are quite concentrated, e.g., “answer question” and “provide description”. Conversely, words in **Oasis** data are more natural and representative.

Oasis data contains diverse language types. Thank to the auto-regressive nature of our method, no language bias will be introduced by text prompt languages. Therefore, the generated instructions cover a wide range of language types. With the help of the langdetect library, we provide a visual breakdown of the language type distribution in **Oasis** and LLaVA-NeXT data, shown in Fig. 3. It is observed that our data contains more than 40 language types, while LLaVA-NeXT data only contains English. To our farthest knowledge, this is the first synthesized multi-modal dataset that covers such a wide range of languages.

Oasis data covers a wide range of root verbs and top noun objects. We provide a visual breakdown of the most common root verbs and top noun objects in Fig. 4. Specifically, we extract the root verbs with a frequency over 1% and their top 3 noun objects using the spaCy library. Compared to the LLaVA-NeXT data, the root verbs in **Oasis** data cover more natural and informative vocabularies. The top

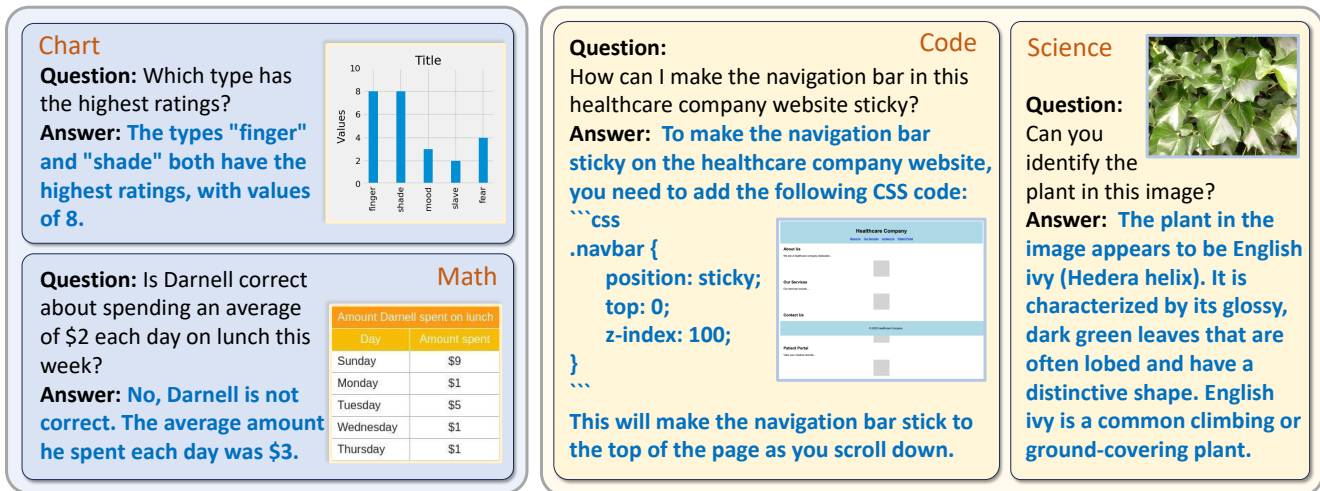


Figure 5. **Oasis synthetic data instances.** We present several examples illustrating the robustness of our data synthesis approach across diverse and rare domains (e.g., Chart, Math, Code and Science). Interestingly, the bottom-left figure shows that **Oasis** can generate complex math questions based on a chart image. **Oasis** identifies patterns and context within the table (e.g., daily spending amounts) and the related textual query ('Is Darnell correct about spending an average of \$2 each day on lunch?')

noun objects are also more diverse. Interestingly, LLaVA-NeXT data contains a lot of the combination “answer question”, while **Oasis** has no such bias. This suggests that **Oasis** captures a broader set of actions and interactions, reflecting a more nuanced and varied set of activities in the dataset. The heavy reliance on “answer questions” in LLaVA-NeXT hints at a potential overemphasis on QA tasks. More word-level analyses are in Appendix A.4.

Cases of Oasis data. We present cases of **Oasis** data in Fig. 5. It should be noted that our method can generate detailed and informative instructions based on the image topic. Additionally, we observe that our synthetic data covers a wide range of tasks, including object recognition, scene description, code understanding, etc. Such illustrations also support the aforementioned arguments about data diversity. More examples could be found in Appendix A.2.

4. Experiments

4.1. Implementation Details

Synthesis details of Oasis-500k data. During the data synthesis process, we utilized two open-source models, Qwen2-VL-72B-Instruct as MLLM and Qwen2.5-72B-Instruct as LLM. We randomly sample images from various public datasets as composed by the Cambrian-10M collection. Above models and dataset are easy to access via Huggingface, which guarantees the reproduction of this work.

Pre-training and supervised-finetuning data. During the pre-training phase, we use LLaVA-Pretrain-558K [23]

for image-text alignment training. For supervised fine-tuning, we use officially open-source LLaVA-NeXT data as the baseline training set, where 15k instruction data from user data are not released. In the main experiment, we add our synthesized **Oasis** data to the original training set to evaluate the effectiveness of our method. In the ablation study, we prove that the multi-stage filtering mechanism improves the data quality.

Model. For reproduction, we strictly follow the architecture of LLaVA-NeXT [23]. The typical multi-modal large model consists of three key components: a visual encoder to extract visual representation, an image-text projector to align the visual and text modalities, and an LLM to generate the answer. We use CLIP-ViT-L as the backbone visual encoder and Vicuna-7B-v1.5 / Qwen2.5-7B-Instruct / Llama3-8B-Instruct as the LLM respectively. The projector is a 2-layer MLP with GELU activation function.

Training recipe. Our experiments are conducted based mainly on the LLaVA-NeXT official codebase. We follow the 2-stage training strategy, including vision-language pre-training and visual instruction-tuning. In the pre-training stage, only the randomly initialized projector is trained. In the fine-tuning stage, the full parameters are tunable. For both pretrain and finetune stages, we use a cosine learning rate schedule with a warmup ratio of 0.03 and a global batch size of 128. The optimizer is AdamW [28] without weight decay. The learning rate is set to 1e-3 for pretrain and 1e-5 for finetune. It takes about 4 hours for pretraining and 30 hours for finetuning the baseline (LLaVA-NeXT).

Table 2. **Benchmarks used in the experiments and their corresponding domains.** We carefully choose 7 domains and 14 benchmarks to evaluate the effectiveness of our method comprehensively.

Domain	Benchmark	Domain	Benchmark	Domain	Benchmark
OCR	TextVQA [36]	Science	ScienceQA [35]	General	MMBench [27], MME [14] MMVet [45], MMstar [6]
	OCRBench [25]		AI2D [16]		
Chart	SeedBench2-Plus [17, 18]	Document	DocVQA [31]	Multi-discipline Visual Reasoning	MMMU-Pro [46] GQA [15]
	ChartQA [30]		InfoVQA [32]		

Benchmarks. To comprehensively evaluate the effectiveness of our evolutionary method, we select 14 benchmarks, with their sources and tested skills illustrated in Tab. 2. These benchmarks encompass a wide range of vision-language tasks, including OCR, chart recognition, document analysis, general knowledge, multi-discipline, and visual reasoning. We evaluate the performance of our method on these benchmarks to demonstrate its effectiveness in improving the generalization ability of multi-modal models.

4.2. Main Results

Effectiveness of Oasis. To demonstrate Oasis’s effectiveness, we perform a thorough evaluation of the MLLM trained with Oasis data across 14 benchmarks. As shown clearly in Fig. 6a, our method provides a comprehensive and substantial improvement over the baseline. The detailed results are presented in Tab. 3. We observe that our method consistently outperforms the baseline in almost all benchmarks, with an average improvement of 3.1% / 1.8% / 3.2% for Vicuna1.5 / Qwen2.5 / Llama3 backbone, respectively. In particular, for Vicuna-7B-v1.5, improvements come from various domains, *e.g.*, general knowledge, OCR, and document analysis. On the general knowledge domain, Oasis achieves 1.4% and 2.3% improvements over the baseline data on MMBench-en/cn. On OCR benchmarks, Oasis increases TextVQA and OCRBench for 2.7% and 2.1%. On document analysis tasks, Oasis gains 4.3% and 6.3% over baseline. The above domains prove the diversity of our synthetic data, and the above improvements reveal the effectiveness of our method in enhancing the generalization ability of MLLMs.

Comparison with other synthesis methods. To further validate the efficiency of our method, we compare Oasis data with four other datasets of equivalent size in Vicuna-7B-v1.5. (1) The original annotation of Oasis images in Cambrian-10M [39]. (2) Upsampled LLaVA SFT data. (3) MMEvol data. (4) DenseFusion-1M data [20] (combined with random detailed description instructions in LLaVA paper). We sample each dataset to the same size with Oasis for fair comparison. The results in Tab. 3 show that Oasis data outperforms other methods remarkably. OCR-related tasks benefit the most from our data, with a

steady improvement of more than 2% across DocVQA, InfoVQA, TextVQA and OCRBench over most methods. On general tasks, Oasis data also shows an overall advantage on MME, MMstar and MM-Vet. We argue that Oasis data is inherently easier for the model to learn thanks to its prompt-free nature, as the generated data adheres to the internal knowledge distribution of MLLMs.

Scaling synthetic data. We conduct a scaling experiment to investigate the impact of the amount of synthetic data on the performance of MLLM. Specifically, we respectively add 150k, 300k, and 500k Oasis data to 100k LLaVA-NeXT SFT data, and compare the performance of these models. We downsample LLaVA-NeXT data to 100k to avoid the common problem of ‘data mixture ratio’ and thus fully reveal our data efficiency, but keep the basic model capacity simultaneously. Table 3 shows that when the size of the data increases from 0 to 500k, the overall performance of the model improves steadily. After incorporating 500K Oasis data, the average score improved by 5.2%, providing strong evidence of the effectiveness of our data. Additionally, it is noteworthy that an increase from 300K to 500K data still results in a substantial 4.0% improvement, indicating that our scaling remains effective even in large data amounts and enables the model to achieve consistent gains. We argue that when scaled up, our synthetic data can continuously inject knowledge into the model, progressively enhancing its capabilities.

4.3. Specific-domain Data Synthesis

Oasis is a flexible method for the synthesis of data in a specific domain, since the input image inherently sets the data domain. By conditioning on the characteristics of the input image, Oasis effectively generates domain-specific attributes, which makes it possible to train models in fields where data might be scarce. We take ‘OCR’ as a typical data domain to validate the effectiveness of Oasis in this section. Additionally, Appendix A.5 includes an application of Oasis in the medical domain.

OCR data source and the synthetic data. We select 311k images from 24 OCR-related datasets, and generate

Table 3. **Main results.** All baseline experiments are reproduced with LLaVA-NeXT’s official code. **Oasis** introduces extra synthetic data to baseline in the fine-tuning part. We observe consistent performance gains across various benchmarks, including general and complex tasks. The improvement **+3.1%** / **+1.8%** / **+3.2%** is particularly notable in average, indicating that the synthetic data effectively supplements real-world data. Our data also outperforms other synthetic data in most benchmarks, with overall improvement ranging from 1.2% to 4.5%. The last part shows the efficiency of scaling **Oasis** data, with 5.2% improvement in average.

Method	MMBench	MME	MMStar	MMVet	MMMUPRO _{std}	GQA	AI2D	Sci	Doc	Info	Chart	Seed2	Text	OCR	AVG.
Baseline-Vicuna1.5	64.2 / 54.4	1482 / 291	37.1	28.0	19.7	63.8	65.2	71.5	71.7	33.3	62.7	50.3	63.4	52.9	53.0
LLaVA (upsample)	64.8 / 54.9	1461 / 353	37.6	34.3	19.8	63.9	65.9	71.6	67.8	29.4	64.4	51.3	64.0	52.6	53.7
Densefusion	67.4 / 56.2	1523 / 333	37.8	30.2	19.4	63.9	65.4	71.9	69.2	32.9	61.5	53.6	65.4	55.4	54.3
Cambrian	66.8 / 56.6	1504 / 329	37.8	32.4	19.7	64.1	69.4	70.6	73.8	37.7	63.4	53.9	63.7	52.3	54.9
MMEvol	63.6 / 53.8	1503 / 316	32.3	34.9	19.1	63.4	64.9	54.4	64.7	30.5	61.5	53.8	62.8	51.7	51.6
Oasis-Vicuna1.5	65.6 / 56.7	1532 / 357	38.0	37.2	19.9	63.5	66.0	72.0	76.0	39.6	65.8	54.5	66.1	55.0	56.1
Baseline-Qwen2.5	76.0 / 74.4	1577 / 405	51.2	32.2	28.7	64.1	76.7	83.1	74.0	35.9	73.1	62.6	65.0	56.0	61.4
Oasis-Qwen2.5	77.4 / 74.5	1598 / 439	51.3	35.2	29.5	64.0	78.7	83.0	77.3	41.6	74.3	65.1	65.9	58.2	63.2
Baseline-Llama3	71.4 / 66.6	1539 / 313	45.6	29.6	21.7	63.8	72.1	80.1	65.8	28.3	64.4	54.8	61.1	51.2	55.8
Oasis-Llama3	72.7 / 67.5	1522 / 355	46.4	36.6	22.4	63.9	74.6	81.0	73.0	34.7	71.2	59.5	64.9	55.8	59.0
LLaVA-100k	57.6 / 46.9	1408 / 284	37.7	26.7	18.6	57.9	56.3	70.8	52.0	25.3	47.3	44.9	56.4	40.0	46.5
+Oasis-150k	55.9 / 48.2	1312 / 297	35.0	29.7	17.7	58.4	56.8	67.7	52.5	25.8	50.2	45.7	57.2	42.4	46.6 _{+0.1}
+Oasis-300k	57.5 / 49.0	1419 / 294	37.3	31.3	18.8	58.0	56.8	68.3	53.8	25.7	51.5	45.4	57.7	44.9	47.7 _{+1.2}
+Oasis-500k	58.8 / 51.4	1448 / 331	40.5	38.3	19.1	58.9	58.0	70.3	63.9	31.7	57.6	52.3	61.9	50.9	51.7 _{+5.2}

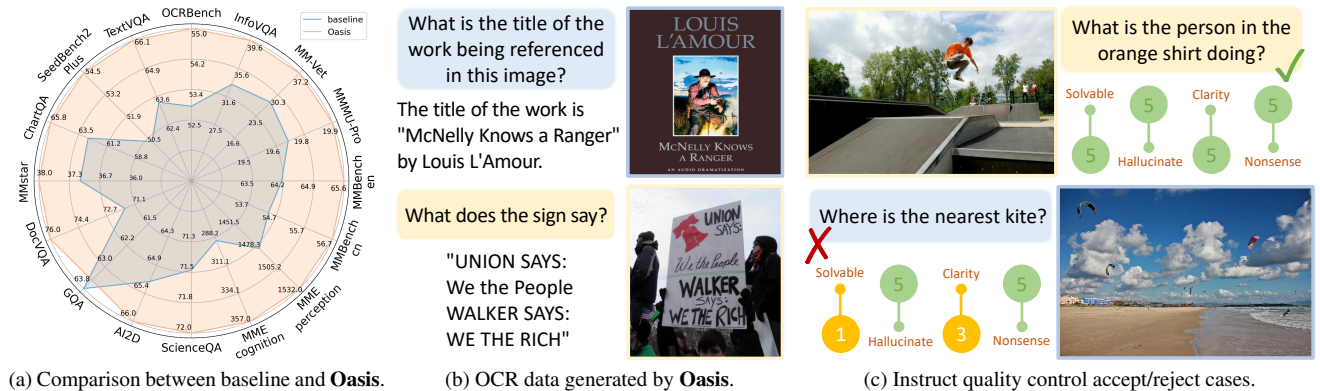


Figure 6. Figure(a) visualizes the improvements of **Oasis** over baseline, and significant overall enhancement can be observed. Figure(b) presents two OCR samples generated by **Oasis**, revealing the data diversity. Figure(c) showcases the mechanism of instruction quality control. We accept the high-quality query above and reject the unsolvable and unclear query in the second case.

Table 4. **OCR domain experiments.** Adding specific-domain data generated by **Oasis** contributes to consistent improvements to various OCR-related benchmarks.

OCR Data	Doc	Text	Chart	OCR	Info	AI2D	Seed2
No	71.7	63.4	62.7	52.9	33.3	65.2	50.3
OCR-30k	73.5	63.7	64.5	54.5	35.6	65.1	52.6
OCR-70k	75.0	64.7	66.2	55.3	37.1	66.3	52.4

data with **Oasis**. Filtered with data categorization and quality control, 70k data remains. We present several examples in Fig. 6b, showing the diversity of **Oasis** data extends beyond basic OCR tasks. It can incorporate a broad spectrum of question types designed to test various reasoning and comprehension abilities. **Oasis** creates questions that not only require direct text extraction but also challenge the model’s capacity for contextual understanding, attribute-

based reasoning, and logical deduction. More discussions and examples can be found in Appendix A.2.

Effectiveness of the synthetic OCR data. As empirical results presented in Tab. 4, the synthetic OCR data demonstrates notable efficacy in improving model performance. We select 7 OCR-related or text-rich benchmarks and find that our OCR-oriented synthesis data could provide steady improvement to these tasks. By introducing additional synthetic data, the model gains exposure to a broader range of problem types, enhancing its ability to generalize to previously unseen tasks. In particular, the synthetic data serves to bridge gaps in the distribution of real-world data, covering edge cases and rare patterns that might otherwise be under-represented. Consequently, this blend of real and synthetic data results in a more robust model that performs reliably across challenging OCR problems.

Table 5. **Ablation results.** The **caption data recycling**, **instruction quality control** and **response quality control** experiments respectively. The caption recycling result provides evidence that we can easily reuse the waste caption data for further improvement. The instruction quality control mechanism resulted in a 1% improvement, proving to be both effective and indispensable. Response quality control methods fail to provide any additional enhancements and are therefore unnecessary.

Method	MMBench	MME	MMStar	MMVet	MMMU ^{pro} _{sid}	GQA	AI2D	Sci	Doc	Info	Chart	Seed2	Text	OCR	AVG.
w/o caption	65.6 / 56.7	1532 / 357	38.0	37.2	19.9	63.5	66.0	72.0	76.0	39.6	65.8	54.5	66.1	55.0	56.1
+ 250k caption	65.8 / 54.6	1496 / 368	40.2	35.9	20.8	64.0	65.7	72.5	77.3	40.5	66.0	55.3	66.3	56.1	56.4
w/ Inst. QC	63.8 / 55.1	1530 / 315	37.9	36.8	19.3	64.0	65.2	71.8	74.8	38.7	64.3	53.7	64.5	52.1	54.9
w/o Inst. QC	64.7 / 52.9	1516 / 308	39.8	35.1	19.0	63.9	65.4	71.4	67.7	31.4	63.8	53.8	63.6	55.1	53.9
w/o Resp. QC	65.6 / 56.7	1532 / 357	38.0	37.2	19.9	63.5	66.0	72.0	76.0	39.6	65.8	54.5	66.1	55.0	56.1
NLL sample	66.1 / 56.2	1535 / 323	39.2	40.3	20.2	64.0	66.3	72.8	69.8	32.8	66.1	53.4	66.0	56.8	55.4
Scoring	56.7 / 54.9	1522 / 288	36.9	43.6	20.5	63.3	65.5	72.6	68.9	32.4	64.5	53.1	64.8	53.7	54.5

4.4. Ablation Studies

Oasis involves two essential synthesis steps, *i.e.*, data categorization and instruction quality control. In the following, we discuss the recycling of the filtered-out data in data categorization and the effects of instruction quality control. Then, we ablate the response quality control is unnecessary.

Recycling of caption data. As stated in Sec. 3.2, we utilize an LLM to process the data obtained from the first step, where the captions are removed while the instructions are retained. Given that the pass rate is only 49.90%, approximately half of the caption-like data remains unused. Therefore, we explore strategies to recycle these data. Since some data contain mixed special tokens or irrelevant fields, we first apply rule-based filtering for an initial screening. Then, we further refine the data using Qwen2.5-72B-Instruct to remove entries that are unsuitable as captions. We apply few-shot in this phase for better accuracy. Eventually, 250k high-quality captions remain. We combine these captions with a random instruction for detailed image description listed in LLaVA paper and add them into the **Oasis**-data. The data in Tab. 5 suggest that this data yields promising results, with 12 out of 16 metrics surpassing the baseline. This result effectively demonstrates that we can efficiently recycle the waste data at an extremely low cost.

Effects of instruction quality control. After data categorization, instruction quality control is applied to filter out low-quality instructions in four dimensions: solvability, clarity, hallucination, and nonsense. Previous works have adequately verified the reliability of LLM/MLLM-as-a-judge for their high human agreement [3, 47]. We further conduct an ablation study to evaluate the impact of this quality control mechanism on the data quality and model performance. In particular, we compare the performance of the model trained with and without quality-controlled 200k data, respectively. Specifically, The acceptance rate of high-quality instructions is a reasonable 50.90%. Therefore, the 200k data without quality control is expected to contain 100k “low-quality” instructions. Table 5 presents the results. With the quality control mech-

anism, the model achieves a notable 1% overall enhancement. In both DocVQA and InfoVQA, the model remarkably gains over 7% improvement. This result demonstrates the necessary role of data quality control in **Oasis**.

Attempts on response quality control. To study the necessity of response quality control, we try 2 methods to filter out low-quality responses. (1) NLL reject sampling: We sample 5 responses for each instruction and calculate their average token NLL (negative log likelihood), and keep the highest one as the final response, which is the answer with the most confidence [33]. (2) MLLM scoring: We score responses from 3 dimensions (helpfulness, truthfulness and instruction-following) on a scale of 1 to 5 with Qwen2-VL-72B-Instruct, and filter out responses that do not get full marks. Results in Tab. 5 reveal that these response quality controls are either ineffective or even harmful. The average score drops by 0.7% and 1.6% in the 2 scenarios. We argue that high-quality instructions could inherently derive good responses from SOTA MLLMs. Unexpected biases would be introduced by extra filtering procedures.

5. Conclusions

Based on the proposed method for synthesizing multi-modal training data, we have demonstrated that it is possible to generate high-quality and diverse multi-modal data using only images, without compromising data quality. Through a careful quality control process and extensive experimentation, we have shown that **Oasis** significantly enhances the performance of MLLM with various backbones, and outperforms other synthesis methods by a large margin. Our approach offers a promising solution to the challenges posed by the unavailability and high cost of traditional multi-modal data synthesis. It also opens new possibilities to improve MLLMs in specific domains. In conclusion, **Oasis** not only addresses the scarcity of multi-modal data but also provides a scalable way to improve the capabilities of MLLM, paving the way for more efficient and accessible training paradigms. We will release our code and dataset to help the future research of the community.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2
- [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. 1, 2
- [3] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024. 8
- [4] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 1, 2
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 2
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 6
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2
- [8] 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. In *SCIC*, 2024.
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2
- [12] Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. Vila²: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024. 2
- [13] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020. 2
- [14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 6
- [15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 6
- [16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 6
- [17] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 6
- [18] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 6
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2
- [20] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and LINGYU DUAN. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. In *NeurIPS*, 2024. 6
- [21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 5
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 1, 2
- [25] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 6
- [26] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024. 2
- [27] 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? In *ECCV*, 2024. 6

- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [29] Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*, 2024. 1, 2
- [30] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 6
- [31] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 6
- [32] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 6
- [33] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *NeurIPS*, 2024. 8
- [34] 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. 2
- [35] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. In *IJDL*, 2022. 6
- [36] 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. 6
- [37] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 2
- [38] Qwen Team. Qwen2.5: A party of foundation models, 2024. 2
- [39] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 4, 6
- [40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2
- [41] 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, 2022. 2
- [42] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *ICLR*, 2024. 2
- [43] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024. 2, 3
- [44] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jizheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2
- [45] 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. 6
- [46] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 6
- [47] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023. 8
- [48] 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. 2