GENIXER: Empowering Multimodal Large Language Model as a Powerful Data Generator

Henry Hengyuan Zhao[1], Pan Zhou[2†], and Mike Zheng Shou[1†]

[1] Show Lab, National University of Singapore, Singapore
[2] Singapore Management University, Singapore

Abstract. Multimodal Large Language Models (MLLMs) demonstrate exceptional problem-solving capabilities, but few research studies aim to gauge the ability to generate visual instruction tuning data. This paper proposes to explore the potential of empowering MLLMs to generate data independently without relying on GPT-4. We introduce GENIXER, a comprehensive data generation pipeline consisting of four key steps: (i) instruction data collection, (ii) instruction template design, (iii) empowering MLLMs, and (iv) data generation and filtering. Additionally, we outline two modes of data generation: task-agnostic and task-specific, enabling controllable output. We demonstrate that a synthetic VQA-like dataset trained with LLaVA1.5 enhances performance on 10 out of 12 multimodal benchmarks. Additionally, the grounding MLLM Shikra, when trained with a REC-like synthetic dataset, shows improvements on 7 out of 8 REC datasets. Through experiments and synthetic data analysis, our findings are: (1) current MLLMs can serve as robust data generators without assistance from GPT-4V; (2) MLLMs trained with task-specific datasets can surpass GPT-4V in generating complex instruction tuning data; (3) synthetic datasets enhance performance across various multimodal benchmarks and help mitigate model hallucinations. The code can be found at https://github.com/zhaohengyuan1/Genixer.

Keywords: Multimodal Large Language Model· Instruction Tuning

1 Introduction

Large Language Models (LLMs) [2, 6, 20, 52] have achieved remarkable success in tackling complex natural language tasks. This progress has recently extended to Multimodal Large Language Models (MLLMs) [3, 10, 17, 20, 21, 25, 34, 40, 55, 63, 74, 78, 81]. These MLLMs exhibit exceptional ability to solve various multimodal problems, but the ability to generate visual instruction data remains an underexplored area. To bridge this gap, this paper is the first to explore the data generation ability of current MLLMs.

Recent MLLMs [3,9,10,14,17,21,39,68] demonstrate that the visual instruction data is essential for multimodal learning. Currently, the data used for model

†Corresponding author

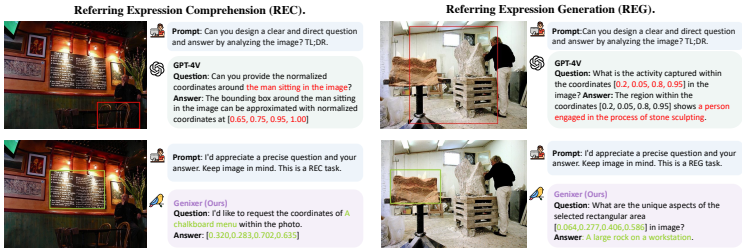


Fig. 1: Two unsatisfied examples generated by GPT-4V [53]. Our proposed data generator GENIXER$_S$ is capable of generating complex multimodal data such as REC and REG data, whereas GPT-4V fails to generate the correct bounding box.

training primarily comes from two methods: (1) reformulating current vision-language datasets into instruction-following format and (2) prompting GPT-4 to create the visual instruction datasets. For the first way, models like InstructBLIP [17] trained on traditional VL tasks, such as visual question answering (VQA) and image captioning, demonstrate effectiveness on downstream tasks. However, the datasets of these tasks suffer from a limitation in image diversity, as most of them originate from the COCO dataset [38], potentially restricting the model generalization ability. For the other way, LLaVA [39], Shikra [10], and ShareGPT4V [11] propose prompting GPT-4 to produce the additional visual instruction tuning data. The primary limitations of prompting GPT-4 are twofold: (1) high financial costs for large-scale dataset creation. (2) inferior performance on some complex tasks such as Referring Expression Comprehension (REC) as illustrated in Fig. 1. Considering these issues, we propose exploring an alternative approach to generating visual instruction tuning data by training an MLLM exclusively for data generation. The benefits of this approach are twofold: (1) it eliminates additional financial costs for data generation, and (2) it provides flexibility in producing high-quality visual instruction data for arbitrary unlabeled images.

In this study, we focus on training an MLLM to learn to generate instruction tuning data from existing datasets, a task we term task-specific data generation. Additionally, we also explore task-agnostic data generation. Task-specific data generation is used to build synthetic datasets tailored to explicit task types, while task-agnostic data generation converts unlabeled images into various task types. To achieve this goal, we introduce a new pipeline GENIXER, as illustrated in Fig. 2. This pipeline consists of four key steps: (i) instruction data collection, (ii) instruction template design, (iii) empowering MLLM, and (iv) data generation and filtering. In the first step, we focus on collecting prevalent and representative vision-language (VL) tasks as sources for data generation. The second step involves meticulously designing two-level instruction templates to enable controlled data generation for both task-specific and task-agnostic. In the third step, we select two representative MLLMs, LLaVA1.5, and Shikra, to cover two main VL task groups data generation: generic and grounding tasks. Both MLLMs are trained to generate instruction data based on their possessed multimodal understanding. In the fourth step, to ensure the quality of the synthetic

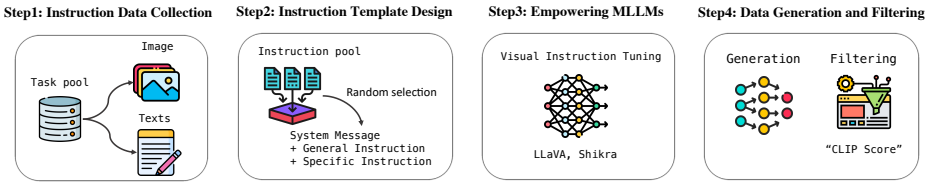


Fig. 2: The illustration of our proposed automatic data generation pipeline GENIXER.

data, we propose removing incorrect data samples using two newly introduced automatic data filtering pipelines.

Using the proposed GENIXER, we build two visual instruction tuning datasets: Genixer-915K for VQA tasks and Genixer-350K for REC tasks. Our experimental results show that training with 915K VQA-like data improves LLaVA1.5 [39] on 10 out of 12 multimodal datasets and benchmarks, such as 3.8% improvement on Vizwiz [23], 2.7% on SicenceQA [46], 37.7% on MME. Additionally, learning with 350K REC-like synthetic data, Shikra shows consistent improvements on 7 out of 8 REC datasets. Beyond quantitative evaluation, human study and dataset analysis, as detailed in Sec. 4.1 and 4.2, demonstrate the diversity and effectiveness of generated instruction tuning data. Visualizing examples of different task types in Fig. 3, it is evident that our GENIXER successfully produces long and meaningful answers on some challenging tasks such as Multi-turn Dialogue, PointQA, and Referential Dialogue.

Overall, our approach yields three main findings: (1) MLLMs trained using our pipeline can produce visual instruction data of comparable quality without the assistance of GPT-4V. (2) MLLMs trained with our pipeline excel at generating more complex instruction tuning data compared to GPT-4V. (3) Synthetic datasets significantly enhance MLLMs' performance on various multimodal benchmarks and help mitigate hallucinations in the models. In summary, our contributions are as follows:

- We present a holistic data generation pipeline, **GENIXER**, capable of generating diverse visual instruction tuning data from unlabeled images.
- We contribute two open-source data generator models **GENIXER**$_L$ and **GENIXER**$_S$ for advancing data creation in multimodal domains.
- We contribute two high-quality multimodal datasets, **Genixer-915K** and **Genixer-350K**, for enhancing other MLLMs in multiple benchmarks.

2 Related Work

Multimodal Large Language Models. Large Language Models (LLMs) showcased the remarkable complicated reasoning abilities. Some well-known opensouced LLMs include FlanT5 [16], OPT [79], LLaMA [65], Vicuna [15] and LLaMA-2 [66] show exceptional reasoning ability of solving math, codes problems. By leveraging these LLMs, a surge of multimodal modes [3, 7–10, 17, 20, 24, 25, 32, 36, 40, 43, 44, 47, 51, 55, 68, 69, 74, 80, 81] are proposed to integrate the

visual information for diverse multimodal reasoning tasks such as image captioning [1, 12, 76] and visual question answering [22, 23, 26, 50]. LLaVA [40] is a pioneering approach that adopts a single linear layer to project the visual features extracted by CLIP [57] encoder to the input of LLM. Different from LLaVA, InstructBLIP [17] employs an instruction-aware feature extractor and obtains advanced performance on various tasks building upon the pretrained BLIP2 [36]. Besides focusing on traditional multimodal tasks, some studies [8,10,58,59,72,75] focus on solving the grounding tasks with the power of LLMs. Shikra [10] and Ferret [8] pay attention to curating the visual instruction data, and PVIT [8] employs the region-based vision encoder. Except for these MLLMs, CogVLM [68], Qwen-VL [3], and Kosmos-2 [55] explore adopting a billion-scale pertaining data corpus to enhance the model generalization ability and robustness.

Multimodal Instruction Data. High-quality visual instruction data is crucial for training an MLLM. Two main setups are as follows: 1) Transforming the current vision-language datasets into the instruction tuning format, e.g., InstructBLIP. Such choice is limited by the diversity of image sources. 2) Some approaches LLaVA, VisionLLM [69], and Shikra, resort to prompting the GPT-4 [52] language model to generate corresponding instruction data. This way requires the image datasets to have enough captions or region-based annotations (e.g., bounding boxes), which heavily restricts the data scale. Additionally, prompting commercial LLMs incurs high costs, and even GPT-4V [53] may not address the data generation effectively on some specific tasks, as illustrated in Fig. 1. Recently, some works [70, 71] in natural language processing domains propose to generate text instruction tuning datasets for performance improvements. In the multimodal domain, some works have also been proposed to address image caption data generation [11] and text-centric instruction tuning data generation [64]. However, both approaches rely on prompting GPT-4V or Gemini Pro to build their synthetic datasets. Unlike these works, we introduce Genixer, an innovative pipeline that explores the capabilities of MLLMs to generate high-quality visual instruction data without assistance from current commercial LLMs. Our approach is the first to demonstrate the effectiveness of training current MLLMs to generate task-specific visual instructing tuning data and improve the MLLMs on several multimodal datasets and benchmarks.

3 Genixer: An Automatic Visual Instruction Tuning Data Generation Pipeline

Though current MLLMs [3, 39, 68] show exceptional capability of handling various multimodal tasks, rare works concentrate on visual instruction tuning data generation. To this end, We propose Genixer, as illustrated in Fig. 2, which is a novel pipeline that contains four key steps, including 1) instruction data collection, 2) instruction template design, 3) empowering MLLMs, and 4) data generation and filtering. In the following, we will elaborate on these four key steps.

Table 1: The statistic of tasks and datasets for training GENIXER. We categorize the VL tasks into two categories: generic and grounding tasks. Counting110K† is built by ourselves derived from PointQA [48]. POPE‡ refers to the object hallucination dataset generated by ourselves via the pipeline provided in POPE [37].

Category	Task	Dataset	Size
Generic	Common VQA	VQAv2, GQA, Counting110K†, POPE‡	583K
	Adv VQA	POPE‡	20K
	MC VQA	A-OKVQA	17K
	MD	VQAv2, LLaVA-Conv-58K	108K
Grounding	REC	VG, RefCOCO	1M
	REG	VG, RefCOCO	1M
	PointQA	PointQA Local, Visual7W	218K
	Q→C^{Box}A	Shikra (GPT-4 Generated)	4K
	RD	Shikra (GPT-4 Generated)	1.8K

3.1 Instruction Data Collection

In accordance with the prevalence and practical relevance of real-world VL tasks, we carefully select nine representative tasks as listed in Tab. 1, including Common Visual Question Answering (Common VQA), Adversarial-based VQA (Adv VQA), Multi-choice VQA (MC VQA), Multi-turn Dialogue (MD), Referring Expression Comprehension (REC), Referring Expression Generation (REG), PointQA, Q→C^{Box}A and Referential Dialogue (RD). We divide these tasks into two main categories, **generic tasks** and **grounding tasks**.

3.2 Instruction Template Design

In an automatic data generation context, where image content is agnostic, preemptively determining the specific task type becomes particularly daunting, especially when it involves large-scale data creation purposes. Hence, we consider two critical modes for visual instruction data generation: 1) task-agnostic and 2) task-specific data generation.

Two-level Instructions. We propose a two-level instruction template for controlling the task type of generated visual instruction tuning data. The instruction template is as follows:

> `<s>` SYSTEM MESSAGE. USER: `<image>` Generic Instruction. Specific Instruction. ASSISTANT: Question: `<question>` Answer: `<answer>` `</s>`

The tags `<image>`, `<question>`, and `<answer>` serve as placeholders for inserting the tokens of the image, question, and answer, respectively. Question: `<question>` Answer: `<answer>` is the model response that needs to be predicted in a left-to-right text generation manner.

Regarding Generic Instruction, it allows the model to generate arbitrary types of instruction tuning data referred to as mode 1. During training, we randomly select one of 58 handwritten instructions each time. For instance, *"Please*

Fig. 3: Selected examples generated from GENIXER$_L$ and GENIXER$_S$. The examples include Common VQA, Adv VQA, MC VQA, MD, and five grounding tasks.

provide a clear and direct question and answer after examining the image". Then, for Specific Instruction, it is designed to determine the specific task type, such as *"This is a Common VQA task"*, allowing us to control the task type referred to as mode 2.

Controlling Constant. During the training phase, we set a constant τ for controlling the ratio of training samples that are exclusive with Generic Instruction. Consequently, in the inference phase, the model is able to switch the mode by adding specific instructions or not. For example, as illustrated in Fig. 4, the model is capable of generating various types of data in the absence of specific instructions. Simultaneously, it can produce specific outputs when guided by a detailed prompt, like *"This is an MC VQA task"*.

3.3 Empowering Current MLLMs

To train an MLLM with the ability of data generation, we leverage the two remarkable MLLMs, LLaVA1.5 and Shikra, as the backbone models for generic

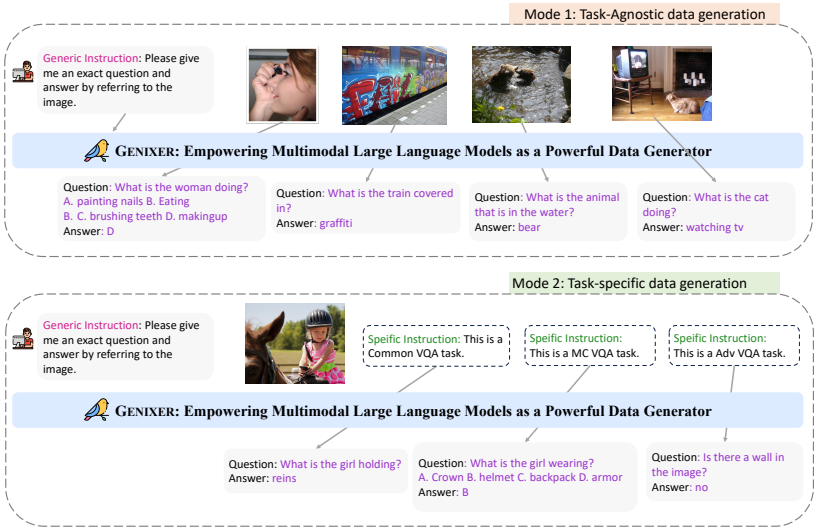


Fig. 4: A demonstration of two proposed instruction modes during the inference phase.

and grounding task generation. Consequently, we obtain two data generators, GENIXER$_L$ and GENIXER$_S$.

Overall Framework of GENIXER$_L$ and GENIXER$_S$. For brevity, we denote the MLLM model as F_M, the generic and specific instructions as X_G and X_S. Then, given an image X_I, training objective is to make F_M to generate question X_q and corresponding answer X_a:

$$X_q, X_a = F_M(X_G, X_S, X_I). \tag{1}$$

To this end, we follow previous MLLMs [17,39], and design the training objective in an autoregressive manner:

$$\max \sum_{i=i}^{L} \log p(X_o|(X_G, X_S, X_I) = \prod_{i}^{L} p_\theta(x_i|(X_G, X_S, X_I, X_{o,<i}), \tag{2}$$

where X_o is the whole sentence compose X_q and X_a, and x_i is current prediction token. L is the length of the model response sequence. θ denotes the total trainable parameters in F_M (e.g., the parameters of projector and LLM with LLaVA1.5 backbone).

Training of GENIXER$_L$. GENIXER$_L$ trains the pretrained LLaVA1.5 [39] for four kinds of generic tasks, including Common VQA, Adv VQA, MC VQA, and MD. As summarized in Tab. 1, we only sample a subset of original datasets from these task types for training efficiency. The controlling constant τ is set to 0.2, 0.2, 0.5, and 0.2, respectively. The different ratios are because of the different data sizes of these task types. These values are chosen manually to keep the training data balanced. Finally, we use AdamW [42] optimizer with a learning rate of 1×10^{-5} and a batch size of 128 for one epoch training takes about 14 hours.

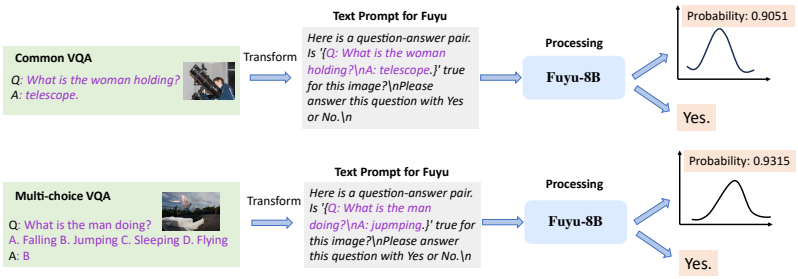


Fig. 5: The illustration of proposed Fuyu-driven data filtering framework. The outputs of the framework compose a probability and a direct answer.

Training of GENIXER$_S$. We utilize the Shikra as the backbone model to train an MLLM with the ability to generate grounding task data. As shown in Tab. 1, the dataset size for RD is relatively small. To counteract potential biases in the generation, GENIXER$_S$ that finetunes a SoTA model Shikra [10] on region-based tasks adopts two-phased training. The first phase focuses on the REC and REG data generation. In the second phase, GENIXER$_S$ adds PointQA, Q→C^{Box}A, and RD data while deliberately reducing REC and REG data for data trade-off. To train the GENIXER$_S$, we set the τ equal to 0.5 for each task type. We utilize the AdamW optimizer, applying learning rates of 3×10^{-5} for the first phase and 1×10^{-5} for the second phase. The batch sizes are set to 128 and 64, respectively.

3.4 Automatic Data Generation and Filtering

Here, we introduce the image sources used for data generation. Additionally, recognizing that there are no ready-made state-of-the-art models available for the VQA data filtering, we introduce two data filtering pipelines to remove the incorrect data samples, which is essential for subsequent model training.

Data Generation and Filtering of GENIXER$_L$ on The Generic Tasks. We utilize the mixed dataset comprising 558K images from the LAION [60], CC3M [61], and SBU [54], as described in [40], and further adopt the 830K images from the original SBU datasets. We directly feed these 1.4M images to the trained GENIXER$_L$ and then designate the specific instruction X_S = *"This is the Common VQA task"*, and finally we produce the 1.4M raw data samples.

To assess the quality of generic task data, we design a Fuyu-driven data filtering framework to automatically filter the samples, which may include incorrect questions or incorrect answers. We design the text prompt as follows:

> Here is a question-answer pair. Is {Q:X_q\nA:X_a} true for this image?\n Please answer this question with Yes or No.\n

For filtering Common VQA and Adv VQA tasks, we substitute the variables X_q and X_a with the generated questions and answers, respectively. In the case of MC VQA evaluation, we replace the option letter (e.g., "B") with the corresponding option content (e.g., "Jumping") and then convert the format to match

that of Common VQA for processing by Fuyu-8B [5], as shown in Fig. 5. For the MD task, we decompose multi-turn dialogue into individual single-turn instances for filtering.

Rather than prompt Fuyu-8B to directly output "Yes" or "No" as the filtering label, we calculate the probability of predicting the "Yes" as follows:

$$P(Y_r|X_I, X_q, X_a) = \prod_{i}^{L} p(y_i|X_I, X_q, X_{a,<i}), \tag{3}$$

where Y_r is the predicted response and L is the length the total response sequence. Then, we propose a threshold λ to control the filtering in the following manner:

$$S^n = \begin{cases} \text{True, if } Y_r = \text{Yes and } P(Y_r^n) > \lambda \\ \text{False, if } Y_r = \text{Yes and } P(Y_r^n) \leq \lambda \\ \text{False, if } Y_r = \text{No,} \end{cases} \tag{4}$$

where S^n is the filter label representing keeping or removing the current sample. $P(Y_r^n)$ denotes the probability of the result "Yes" of n-th candidates. By setting $\lambda = 0.7$, we filter the 1.4M raw VQA triplets to 915K instances. We name this VQA-like synthetic dataset **Genixer-915K**.

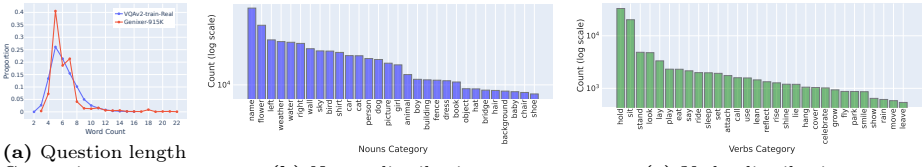
Data Generation and Filtering of GENIXER$_S$ on The Grounding Tasks. GENIXER$_S$ adopts the same image resources in generic tasks for generating grounding-based instruction tuning data. After feeding the 1.4M image corpus to GENIXER$_S$ by set the data type as REC, we get 1.4M raw data.

To assess the quality of these REC data, we propose a CLIP-driven data filtering framework. Specifically, we first use the regular expression to extract the text expression and region coordinates from the raw generated sentence and then conduct the following three steps to filter the generated data in a coarse-to-fine manner. 1) Removing the wrong question or answer formats (e.g., wrong coordinate format). 2) Removing the bonding box whose width or height is smaller than 50. 3) Employing OpenCLIP-L [27] model for calculating the similarity score between the text expression and their corresponding image region, discarding samples with CLIP scores below 0.6. For example, consider one REC sample with the Question *"I need the coordinates of the person at the bottom left of the image. Can you assist?"* and the Answer *"[0.005,0.332, 0.249,0.984]"*. Here, *"person at the bottom left of the image"* is the text expression, and the coordinate of the referring region is *"[0.005,0.332,0.249,0.984]"*.

By applying a threshold of 0.6, we filter out 350K instances from 1.4M images and name this REC-like synthetic dataset as **Genixer-350K**.

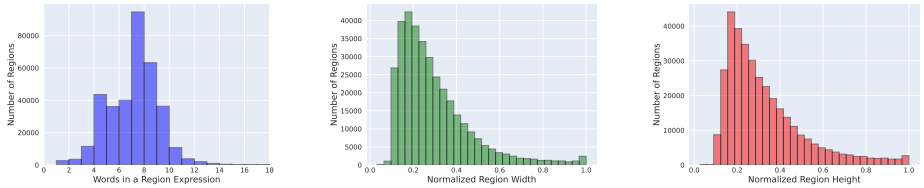
4 Experiments

In this section, we evaluate the quality of synthetic datasets from several aspects, including statistical analysis, human evaluation, evaluation via training MLLMs, ablation study.



(a) Question length Comparison. **(b)** Nouns distribution. **(c)** Verbs distribution.

Fig. 6: The statistics of VQA-like dataset Genixer-915K.



(a) Expression Length. **(b)** Region width. **(c)** Region Height.

Fig. 7: The statistics of REC-like dataset Genixer-350K.

4.1 Statistical Analysis

To evaluate the generation quality of GENIXER$_L$, we conduct a comparative analysis using the VQAv2 [22] training set as a comparison. Fig. 6 (a) illustrates the distribution of question lengths of VQAv2 and **Genixer-915K**. Notably, our dataset exhibits a significant long tail, indicating a higher proportion of longer sentences compared to VQAv2. Additionally, the distribution of nouns and verbs, depicted in Fig. 6 (b) and (c), showcases the diverse vocabulary present in the generated questions. To assess data quality, we utilize Flickr30K [76] as the image source and employ GENIXER$_L$ to generate three representative data types, as detailed in Sec. 3.4. Results in Tab. 2 demonstrate that our generated data achieves an accuracy exceeding 80%, with the highest accuracy observed in the MC VQA category. Furthermore, our generated data exhibits a high probability (0.8) of being classified as "Yes" across all three data types. This is corroborated by the probability distribution depicted in Fig. 8, affirming the high quality of data produced by GENIXER$_L$ in generating diverse instruction tuning data.

The statistics of the synthetic dataset **Genixer-350K** are presented in Fig. 7, showcasing metrics related to expression length, region width, and height. Tab. 3 offers a comparative analysis of our dataset, highlighting the larger collection of images and expression lengths in Genixer-350K compared to other grounding-based datasets.

4.2 Human Evaluation

We conduct the human evaluation to manually analyze the generated question type and corresponding correctness. We employ GENIXER$_L$ to generate QA samples without specific the data type. Since the image content can affect the generation types, we randomly chose 100 images from the COCO validation set as

Table 2: Fuyu-8B evaluation result on Flickr30K image dataset. Accuracy refers to the "Yes" prediction.

Data Type	Accuracy(∼%)	Average Prob.
Common VQA	82.4	0.8186
MC VQA	87.8	0.8721
MD	82.5	0.8252

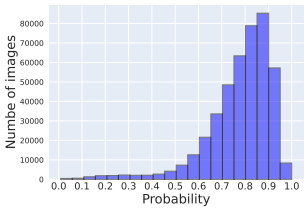


Fig. 8: The distribution of the probability by Fuyu-8B evaluation on Genixer-915K.

Table 3: Comparison of images, objects, and average length between Genixer-350K with other visual grounding datasets.

Dataset	Images	Objects	Avg. Length
Flickr Entities [56]	31,783	275,775	–
RefCOCOg [49]	26,711	54,822	8.43
RefCOCO [30]	19,994	50,000	3.61
RefCOCO+ [30]	19,992	49,856	3.53
Visual Genome [31]	108,077	4,102,818	–
Genixer-350K	350,000	447,801	6.67

Table 4: The human evaluation on 100 randomly selected examples from held-in (COCO Val) and held-out (Flickr30K) datasets.

Question Type	Held-in (COCO val)		Held-out (Flickr30K)	
	#Samples	Correct (∼%)	#Samples	Correct(∼%)
Action	13	92	17	88
Color	8	75	4	75
Counting	6	83	3	66
Object Type	23	87	38	76
Relative Position	32	75	23	65
Yes/No	2	50	4	100
Others	16	81	11	82

the held-in set and 100 images from Flickr30K as the held-out set. As shown in Tab. 4, we divide the questions into seven types, such as "Action" and "Color". One can observe that "Object Type" and "Relative position" are the two most popular question types of both held-in and held-out datasets. Among held-in samples, "Action" question also has a notable proportion with 92% correctness. Different from the held-in set, 38 out of 100 samples belong to "Object Type" and 23 out of 100 "Relative Position" in the held-out set. In all of these types, the correctness of the held-in dataset is slightly higher than the held-out dataset.

4.3 Evaluation via Training MLLMs

Unless specifically stated otherwise, all experiments were conducted on an 8 Nvidia A100 (40G) GPU setup.

Evaluation on General Tasks.

Benchmarks. Following the baseline LLaVA1.5 [39], we evaluate the enhanced model on 12 multimodal datasets and benchmarks. We select five generic VQA datasets, including VQAv2 [22], GQA [26], VizWiz [23], ScienceQA [46], and TextVQA [62]. We test the multimodal benchmarks on MME [18], MMBench [41], SEED-Bench [35], LLaVA-Bench [40] and MM-Vet [77]. Additionally, we also test the model on hallucination benchmark POPE [37].

Main results. As pointed out in [17] and [39], the ratio of dataset mixture is crucial for the model training. We proceed to add the large-scale synthetic dataset Genixer-915K into the pretraining stage rather than the finetuning stage for a

Table 5: Comparison with SoTA methods on 12 benchmarks. * represents the train set used in training. All abbreviated names of benchmarks are following [39]. † indicates results we reproduced since the original results could not be replicated.

Method	LLM	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	$SEED^{I}$	$LLaVA^{W}$	MM-Vet
BLIP-2 [36]	Vicuna-13B	41.0	41.3	19.6	61.0	42.5	85.3	1293.8	–	–	49.7	38.1	22.4
InstructBLIP [17]	Vicuna-7B	–	49.2	34.5	60.5	50.1	–	–	36.0	23.7	58.8	60.9	26.2
InstructBLIP [17]	Vicuna-13B	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
Shikra [10]	Vicuna-13B	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
IDEFICS-9B [33]	LLaMA-7B	50.9	38.4	35.5	44.2	25.9	–	–	48.2	25.2	44.5	–	–
IDEFICS-80B [33]	LLaMA-65B	60.0	45.2	36.0	68.9	30.9	–	–	54.5	38.1	53.2	–	–
Qwen-VL [3]	Qwen-7B	78.8*	59.3*	35.2	67.1	**63.8**	–	–	38.2	7.4	62.3	–	–
Qwen-VL-Chat [3]	Qwen-7B	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	60.6	56.7	65.4	–	–
LLaVA-1.5	Vicuna-7B	78.5*	62.0*	50.0	66.8	58.2	85.9	1465.0†	64.3	58.3	66.2	**65.4**	**31.1**
LLaVA-1.5$_{+G\text{-}910K(ours)}$	Vicuna-7B	**79.1***	**63.1***	**53.8**	**69.7**	59.0	**87.3**	**1502.7**	**65.3**	**59.4**	**66.6**	64.0	30.1

Table 6: Results on Referring Expression Comprehension (REC) task.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg
	val	test-A	test-B	val	test-A	test-B	val	test	
GPV-2 [29]	51.59	–	–	–	–	–	–	–	–
OFA-L [67]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
Unified-IO [45]	78.60	–	–	–	–	–	–	–	–
OFASys [4]	–	80.10	–	–	–	–	–	–	–
VisionLLM-H [69]	–	86.70	–	–	–	–	–	–	–
UNITER [13]	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	76.17
VILLA [19]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	77.77
UniTAB [73]	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	80.89
MDETR [28]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	81.81
Shikra [10]	87.01	90.61	80.24	81.60	87.36	72.12	**82.27**	82.19	82.92
Shikra$_{+G\text{-}350K(ours)}$	**87.48**	**91.05**	**81.77**	**81.89**	**87.43**	**73.14**	81.99	**83.15**	**83.49**

fair comparison, adhering to the same training protocols utilized by LLaVA1.5. For the finetuning stage, we meticulously curated 8K VQA-like data samples, selected based on Fuyu-8B's probability range of 0.5 to 0.7 with the image resource from SBU [54]. The probability range was chosen based on findings that higher probabilities correlate with simpler, more straightforward VQA instances. By selecting data within this specific range, our model can learn more challenging VQA samples. From Tab. 5, one can observe consistent enhancements across 10 out of 12 benchmarks compared with vanilla LLaVA1.5. Indeed, on several tasks, our generated data can make significant improvements, e.g., 3.8% on VizWiz, 2.9% on SciencQA, and 37.7 scores on the MME benchmark. While the improvement in the SEED benchmark is not as obvious as we initially expected, this is primarily because the synthetic data belongs to the Common VQA type while many questions of SEED require outside knowledge to solve which is out of the scope of Genixer-915K. The performance on $LLaVA\text{-}Bench^{W}$ and MM-Vet shows a slight decline, possibly due to their smaller scale and the utilization of GPT-4 as the evaluation metric. This could introduce additional uncertainty and potential biases into the results.

Evaluation on Grounding Tasks.

Table 7: Performance of GENIXER$_L$ on 6 representative benchmarks.

Method	Setting	VQAv2	GQA	VizWiz	SQA^I	POPE	SEED^I
LLaVA1.5	-	78.5	62.0	50.0	66.8	85.9	66.2
GENIXER$_L$	ZS	$79.7_{+1.2}$	$61.3_{-0.7}$	$50.8_{+0.8}$	$65.6_{-1.2}$	$86.0_{+0.1}$	$65.6_{-0.6}$
GENIXER$_L$	MixT	$80.2_{+1.7}$	$63.1_{+1.1}$	$54.1_{+4.1}$	$67.1_{+0.3}$	$87.5_{+1.6}$	$67.1_{+0.9}$

Table 8: The effect of Data scales on synthetic VQA-like dataset.

Dataset	VQAv2	GQA	VizWiz	SQA^I	POPE	SEED^I
Baseline	78.5	62.0	50.0	66.8	85.9	66.2
Genxier-300K	79.0	62.9	52.7	68.5	87.1	65.8
Genxier-610K	79.0	63.1	53.7	69.2	87.2	66.2
Genixer-915K	**79.1**	**63.1**	**53.8**	**69.7**	**87.3**	**66.6**

Table 9: The effect of different probability threshold λ.

Setting	Size	VQAv2	GQA	VizWiz	SQA^I	POPE	SEED^I
Baseline	-	78.5	62.0	50.0	66.8	85.9	66.2
$\lambda = 0$	1.4M	79.0	62.9	53.5	69.6	87.1	66.2
$\lambda = 0.5$	1.1M	79.1	63.1	53.2	69.1	86.9	66.4
$\lambda = 0.7$	0.9M	**79.1**	**63.1**	**53.8**	**69.7**	**87.3**	**66.6**

Benchmarks. Following the baseline model Shikra [10], we test the enhanced model on REC tasks. We use the test datasets RefCOCO [30], RefCOCO+ [30], RefCOCOg [49].

Main results. Here, we adopt Shikra [10] as the model to evaluate the quality of Genixer-350K produced by GENIXER$_S$. Adhering to Shikra's published code, we incorporate our synthetic data into the training phases, maintaining consistent training iterations to ensure a fair comparison. Tab. 6 shows the improvement on 7 out of 8 test datasets with a non-trivial average boost of 0.6%. These findings imply that our pipeline can be an alternative approach to generate grounding-based instruction tuning data, which is typically challenging for manually labeling and not satisfied for prompting GPT-4V, as shown in Fig. 1.

Performance on Genixer. As illustrated in Tab. 2 and Fig. 8, GENIXER$_L$ showcases a superior capability of generating high-quality data. It is natural for us to investigate the performance of GENIXER$_L$. Accordingly, we evaluate GENIXER$_L$ and report its results in Tab. 7, which refers to the setting ZS. One can observe the minor declines in performance on the GQA, ScienceQA, and SEED, alongside a modest enhancement on VQAv2, VizWiz, and POPE. Such results are due to the exclusive training on generating instruction tuning data. Thus, for a fair comparison to investigate the capability of GENIXER$_L$, we proceeded to retrain the GENIXER$_L$ with the mixture of the 665K finetuning dataset used in LLaVA1.5 and the datasets for training GENIXER$_L$ for one epoch following the same training protocols. The outcomes of this process are presented as MixT in Tab. 7, where we witnessed significant improvements across all six benchmarks.

4.4 Ablation Study

Effect of data scale. Tab. 8 investigates the effects of the scales of our synthetic data in the pretraining stage. One can observe that a larger scale often leads to a steady performance improvement on all of the six benchmarks, showing the quality of our synthetic data.

Probability Range. Tab. 9 investigates the impact of the probability threshold during data filtering in Section 3.4 on the data quality. By varying λ, we observe that higher values of λ often better improve performance across all six benchmarks, even with a reduced number of selected training samples. This suggests that the quality of data is more crucial than the quantity of samples.

5 Conclusion, Limitations, and Societal Impacts

In this paper, we introduce a novel automatic data generation pipeline called GENIXER, designed to efficiently and affordably produce high-quality instruction tuning data by leveraging current MLLMs. We instantiate GENIXER into two data-generative MLLMs, GENIXER$_L$ and GENIXER$_S$, tailored to generate general and grounding instruction tuning data, respectively. To ensure the quality of the generated data, we propose two data filtering frameworks: Fuyu-driven and CLIP-driven. Finally, we contribute two instruction tuning datasets, Genixer-915K and Genixer-350K, targeting Common VQA and REC. Experimental results demonstrate that both generated datasets significantly enhance LLaVA1.5 and Shikra across various multimodal benchmarks, respectively.

Limitations. *1) LLM Scale:* Due to computational constraints, we do not test larger LLM models, such as 13B or 34B. However, we believe that our data generator could be beneficial since larger models are more data-hungry. *2) Data Scale:* While scaling up the candidate image corpus to larger datasets like LAION-2B could enhance the model capability, training costs and time constraints restrict us to do such expansions. But Tab. 8 shows scaling can improve performance. *3) Evaluation:* Despite proposing effective data filtering frameworks, evaluating complex and open-ended data types like Referential Dialogue remains challenging, leaving room for future exploration.

Societal Impacts. Our work addresses the challenge of generating high-quality instruction tuning data by presenting a comprehensive pipeline. It paves the way for future investigations into generating diverse multimodal data, contributing to advancements in various fields.

6 Acknowledgement

This research is supported by National Research Foundation, Singapore and A*STAR, under its RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) grant call (Grant No. I2001E0059) – SIA-NUS Digital Aviation Corp Lab. Mike Zheng Shou is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. Pan Zhou was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV (2019) 4
2. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv:2305.10403 (2023) 1
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 1, 3, 4, 12
4. Bai, J., Men, R., Yang, H., Ren, X., Dang, K., Zhang, Y., Zhou, X., Wang, P., Tan, S., Yang, A., et al.: Ofasys: A multi-modal multi-task learning system for building generalist models. arXiv:2212.04408 (2022) 12
5. Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., Taşırlar, S.: Introducing our multimodal models (2023), https://www.adept.ai/blog/fuyu-8b 9
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020) 1
7. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced projector for multimodal llm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
8. Chen, C., Qin, R., Luo, F., Mi, X., Li, P., Sun, M., Liu, Y.: Position-enhanced visual instruction tuning for multimodal large language models. arXiv preprint arXiv:2308.13437 (2023) 3, 4
9. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 1, 3
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv:2306.15195 (2023) 1, 2, 3, 4, 8, 12, 13
11. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 2, 4
12. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325 (2015) 4
13. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020) 12
14. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 1
15. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 3

16. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 3
17. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. NeurIPS (2023) 1, 2, 3, 4, 7, 11, 12
18. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394 (2023) 11
19. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. Advances in Neural Information Processing Systems **33**, 6616–6628 (2020) 12
20. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv:2304.15010 (2023) 1, 3
21. Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024) 1
22. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017) 4, 10, 11
23. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018) 3, 4, 11
24. He, M., Liu, Y., Wu, B., Yuan, J., Wang, Y., Huang, T., Zhao, B.: Efficient multimodal learning from data-centric perspective. arXiv preprint arXiv:2402.11530 (2024) 3
25. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv:2302.14045 (2023) 1, 3
26. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019) 4, 11
27. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (2021), https://doi.org/10.5281/zenodo.5143773 9
28. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 12
29. Kamath, A., Clark, C., Gupta, T., Kolve, E., Hoiem, D., Kembhavi, A.: Webly supervised concept expansion for general purpose vision models. In: European Conference on Computer Vision. pp. 662–681. Springer (2022) 12
30. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 11, 13
31. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 11
32. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 3

33. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023) 12
34. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv:2305.03726 (2023) 1
35. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension (2023) 11
36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597 (2023) 3, 4, 12
37. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv:2305.10355 (2023) 5, 11
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 2
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 1, 2, 3, 4, 7, 11, 12
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 1, 3, 4, 8, 11
41. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023) 11
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7 7
43. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Sun, Y., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024) 3
44. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172 (2023) 3
45. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv:2206.08916 (2022) 12
46. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. NeurIPS (2022) 3, 11
47. Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models. NeurIPS (2023) 3
48. Mani, A., Yoo, N., Hinthorn, W., Russakovsky, O.: Point and ask: Incorporating pointing into visual question answering. arXiv preprint arXiv:2011.13681 (2020) 5
49. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 11, 13
50. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019) 4
51. Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.F., Murugesan, P., Heidari, P., Liu, Y., et al.: Anymal: An efficient and scalable any-modality augmented language model. arXiv preprint arXiv:2309.16058 (2023) 3
52. OpenAI: Gpt-4 technical report (2023) 1, 4
53. OpenAI: Gpt-4v(ision) system card (2023), https://cdn.openai.com/papers/GPTV_System_Card.pdf 2, 4

54. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. NeurIPS (2011) 8, 12
55. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv:2306.14824 (2023) 1, 3, 4
56. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision **123**, 74–93 (2015) 11
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4
58. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4
59. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model (2023) 4
60. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 (2022) 8
61. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 8
62. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 11
63. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv:2305.16355 (2023) 1
64. Tang, J., Lin, C., Zhao, Z., Wei, S., Wu, B., Liu, Q., Feng, H., Li, Y., Wang, S., Liao, L., et al.: Textsquare: Scaling up text-centric visual instruction tuning. arXiv preprint arXiv:2404.12803 (2024) 4
65. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023) 3
66. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 3
67. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICML (2022) 12
68. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models. arXiv (2023) 1, 3, 4
69. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023) 3, 4, 12

70. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13484–13508. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.754, https://aclanthology.org/2023.acl-long.754 4
71. Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Lin, Q., Jiang, D.: WizardLM: Empowering large pre-trained language models to follow complex instructions. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=CfXh93NDgH 4
72. Xu, J., Zhou, X., Yan, S., Gu, X., Arnab, A., Sun, C., Wang, X., Schmid, C.: Pixel Aligned Language Models. arXiv preprint arXiv: 2312.09237 (2023) 4
73. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision. pp. 521–539. Springer (2022) 12
74. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv:2304.14178 (2023) 1, 3
75. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2msbbX3ydD 4
76. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. ACL (2014) 4, 10
77. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: International conference on machine learning. PMLR (2024) 11
78. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv:2303.16199 (2023) 1
79. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv:2205.01068 (2022) 3
80. Zhao, H.H., Zhou, P., Gao, D., Shou, M.Z.: Lova3: Learning to visual question answering, asking and assessment (2024) 3
81. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592 (2023) 1, 3