

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

Henry Hengyuan Zhao<sup>1</sup>, Pan Zhou<sup>2,3†</sup>, and Mike Zheng Shou<sup>1†</sup>

<sup>1</sup> Show Lab, National University of Singapore, Singapore

<sup>2</sup> Singapore Management University, Singapore

<sup>3</sup> Sea AI Lab, Singapore

**Abstract.** Instruction tuning data is essential for training the Multimodal Large Language Models (MLLMs). However, the creation of high-quality instruction tuning data presents significant challenges. Asking the human to label the instruction tuning data is label-intensive and time-consuming. Some works prompted to GPT-4 for data generation were not only costly but also lacked satisfactory performance in complex tasks (i.e., grounding-based reasoning tasks). To address the challenges of data creation, we are the first to explore the potential of empowering MLLMs with the ability to generate instruction-tuning data by following user instructions. Specifically, we developed an innovative data generation pipeline **GENIXER** to generate various high-quality instruction tuning data, including nine representative tasks, e.g., Common VQA, REC, REG, and PointQ. GENIXER provides a unified solution for data generation with four key steps: (i) instruction data collection, (ii) instruction template design, (iii) empowering MLLM, and (iv) data generation and filtering. To validate the effectiveness of generated data, we conducted the human evaluation and user preference study to assess the quality of generated data. Subsequently, we generated two instruction-tuning datasets for the training of two representative MLLMs, LLaVA1.5 and Shikra, and noted consistent improvements across various VQA tasks and multimodal benchmarks. For instance, performance on the VizWiz benchmark improved from 50.0% to 53.8%, and on ScienceQA, it increased from 66.8% to 69.7%, reconfirming the quality of the generated instruction tuning data. The data, code, and models will be released.

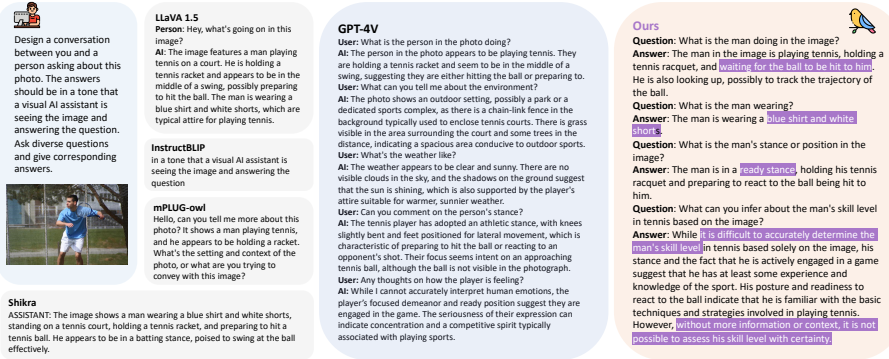
**Keywords:** Multimodal Large Language Model · Instruction Tuning

## 1 Introduction

Large Language Models (LLMs) [2, 5, 15, 39] have achieved remarkable success in understanding human instructions and demonstrated their high proficiency in tackling complex natural language tasks. This progress has recently extended to Multimodal Large Language Models (MLLMs) [3, 8, 13, 15, 18, 25, 31, 42, 47,

---

<sup>†</sup>Corresponding author

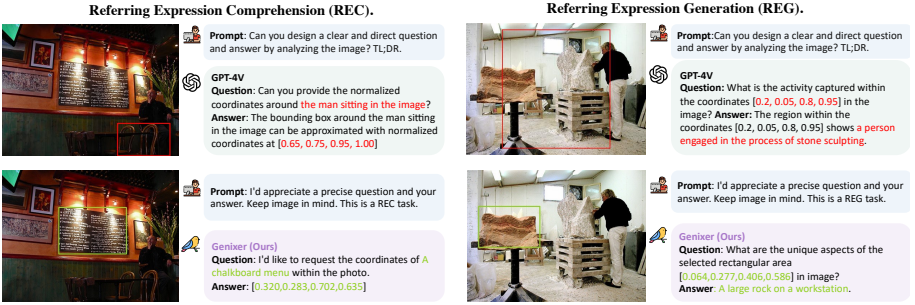


**Fig. 1:** Comparison of prompting the current MLLMs to generate a conversation. LLaVA1.5 [30], mPLUG-owl [54], and Shikra [8] treat the conversation generation task as a dense captioning task while InstructBLIP [13] directly fails. Our GENIXER<sub>L</sub> can generate the question and answer pairs versatily and intelligently and even performs similarly to GPT-4V, as highlighted in purple.

[54, 56, 58], where visual information is encoded by a dedicated visual encoder and seamlessly integrated into LLMs. This integration empowers MLLMs with the ability to analyze and reason about multimodal tasks effectively.

The training of a satisfactory MLLM heavily relies on the availability of sufficient high-quality visual instruction tuning data. The data are, in many ways, the lifeblood of contemporary general-purpose intelligence systems, much like oil powers industrial processes. To attain superior performance, one pioneer model like InstructBLIP [13] requires training on multiple held-in datasets, such as VQAv2 [16] and OKVQA [37], from traditional Vision-Language (VL) tasks. Unfortunately, these held-in datasets suffer from a limitation in image diversity, as most of them originate from the COCO dataset [29], potentially restricting the model generalization ability on unseen visual data. This naturally raises the fundamental question: how can we obtain a sufficient quantity of high-quality instruction tuning data for training MLLMs?

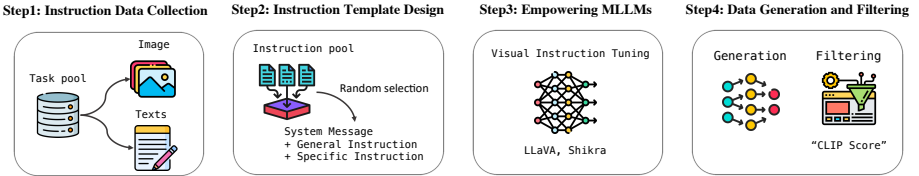
As of now, there is no effective solution to this challenge yet. The traditional approach of manual labeling proves prohibitively expensive and impractical, particularly since MLLM training often necessitates million-scale instruction tuning data to ensure robust generalization. Another potential solution is to leverage powerful models to generate high-quality data. Unfortunately, almost all open-sourced pretrained MLLMs fall short in this regard. As exemplified in Fig. 1, models like LLaVA-1.5 [30], mPLUG-owl [54], and Shikra [8] tend to treat conversation generation as a dense captioning task and struggle to produce natural and coherent dialog data and InstructBLIP [13] directly fails. Contrarily, our produced dialogue exhibits both versatility and intelligence and is comparable to GPT-4V [40]. Except for leveraging current MLLMs, another viable way is to utilize GPT-4V with meticulously designed prompts via the released API, but such a choice leads to high financial costs. For instance, producing millions of samples, each with thousands of tokens, with GPT-4V [40] would cost millions. Moreover, as illustrated in Fig. 2, GPT-4V cannot



**Fig. 2:** Two unsatisfied generation examples from GPT-4V [40]. Our proposed data generator GENIXER<sub>S</sub> is capable of generating complex multimodal data such as REC and REG data, whereas GPT-4V fails to accurately generate the correct bounding box.

generate satisfied instruction tuning data of some complex data types (e.g., referring expression comprehension), further limiting its usage for exclusive model learning.

In this work, we address the challenges of generating highly qualified multimodal instruction tuning data. To this end, we introduce GENIXER, an automatic data generation pipeline that provides a unified solution with four key steps: (i) instruction data collection, (ii) instruction template design, (iii) empowering MLLM, and (iv) data generation and filtering. We begin by categorizing current Vision-Language (VL) tasks into general and grounding tasks. These categories include nine data types, e.g., Common VQA, Multi-choice VQA, Referring Expression Comprehension (REC), and Referential Dialogue (RD). See details in Sec. 3.1. Then, we compile datasets corresponding to these nine tasks and consistently reformat the input data into the instruction-following format. To acquire the flexibility of automatic large-scale and user-guided data generation, we propose a two-level instruction mechanism for creating data in two modes: type-agnostic generation and type-specific generation (see Sec. 3.2). In the type-agnostic mode, given one image, users can pose an instruction without specifying a particular data type to allow the model to perform data generation freely. This is because, in some cases, preemptively determining the optimal data type for generation becomes particularly daunting, especially when it involves large-scale data production for model training purposes. For the type-specific mode, users can ask the model to generate a specific data type from the above nine data types. Subsequently, we leverage two representative models, LLaVA1.5 [30] and Shikra [8] evolve them into the intelligent data generators with the abovementioned two-level instruction template yielding two data generators, GENIXER<sub>L</sub> and GENIXER<sub>S</sub>, for general and grounding data generation, respectively (See Sec. 3.3). To maintain high data quality, we employ the third-party MLLM Fuyu-8B [4] as an evaluator within our specially designed evaluation framework for general task evaluation, detailed in Sec. 3.4. Similarly, we utilize Open-CLIP [20] to calculate the



**Fig. 3:** The illustration of our proposed automatic data generation pipeline GENIXER.

region-expression similarity for REC-like data filtering, enabling us to filter out samples exhibiting low similarity.

Experimental results confirm that using 915K VQA-like tuning data generated by  $\text{GENIXER}_L$  to train LLaVA1.5 can make a significant performance enhancement compared to the vanilla LLaVA1.5 across a diverse set of tasks, such as 3.8% improvement on Vizwiz [17] and 2.7% on SienceQA [33]. Furthermore, our  $\text{GENIXER}_S$  extends its utility to generate instruction-tuning data for visual grounding tasks akin to those encountered in the SoTA grounding model, Shikra. These results underscore the remarkable efficacy of our proposed data generation pipeline, GENIXER. In summary, our contributions are:

- We introduce an innovative multimodal data generation pipeline, **GENIXER**, that enables existing MLLM models to function as potent data generators to effectively and affordably generate high-quality instruction data.
- We contribute two data generators **GENIXER<sub>L</sub>** and **GENIXER<sub>S</sub>** for a wide range of multimodal instruction tuning data generation.
- We contribute two high-quality multimodal datasets: 915K VQA-like data, which can greatly improve LLaVA1.5 on various tasks like Vizwiz and ScienceQA, and 350K REC-like data that enhance Shikra on the grounding tasks like REC datasets.

## 2 Related Work

**Multimodal Large Language Models.** Large Language Models (LLMs) showcased the remarkable complicated reasoning abilities. Some well-known open-sourced LLMs include FlanT5 [12], OPT [57], LLaMA [48], Vicuna [11] and LLaMA-2 [49] show exceptional reasoning ability of solving math, codes problems. By leveraging these LLMs, a surge of multimodal modes [3, 6–8, 13, 15, 18, 27, 31, 34, 38, 42, 51, 52, 54, 58] are proposed to integrate the visual information for diverse multimodal reasoning tasks such as image captioning [1, 9, 55] and visual question answering [16, 17, 19, 37]. LLaVA [31] is a pioneering approach that adopts a single linear layer to project the CLIP [44] encoder extracted visual features to the input of LLM. Different from LLaVA, InstructBLIP [13] employs an instruction-aware feature extractor and obtains advanced performance on various tasks building upon the pretrained BLIP2 [27]. Besides focusing on traditional multimodal tasks, Shikra [8] and PVIT [6] pay attention to the grounding multimodal tasks by carefully designing the visual instruction data or employing the region-based vision encoder. Except for these MLLMs,



**Table 1:** The details of GENIXER training data. We categorize the VL tasks into two categories: general tasks and grounding tasks. Counting110K<sup>†</sup> is built by ourselves derived from PointQA [35]. POPE<sup>‡</sup> refers to the object hallucination dataset generated by ourselves via the pipeline provided in POPE [28].

Category	Task	Dataset	Size
General	Common VQA	VQAv2, GQA, Counting110K <sup>†</sup> , POPE <sup>‡</sup>	583K
	Adv VQA	POPE <sup>‡</sup>	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 <sup>Box</sup> A	Shikra (GPT-4 Generated)	4K
	RD	Shikra (GPT-4 Generated)	1.8K

CogVLM [51], Qwen-VL [3], and Kosmos-2 [42] explore adopting a billion-scale pertaining data corpus to enhance the model generalization ability and robustness.

**Multimodal Instruction Data.** High-quality multimodal instruction data is crucial for training a robust 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 [52], and Shikra, resort to prompting the GPT-4 [39] 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 [40] may not address the data generation effectively on some specific tasks, as illustrated in Fig. 2. To this end, we introduce GENIXER, an innovative pipeline that explores the capabilities of MLLMs to generate high-quality multimodal instruction data.

### 3 GENIXER: An Automatic Instruction Tuning Data Generation Pipeline

Though current MLLMs show exceptional capability of handling various multimodal tasks [3, 26, 30], rare works are concentrating on exploring the data generation capability. In this work, instead of focusing on how to make an MLLM a good problem-solver, we try to provoke the potential ability of current MLLM to generate instruction tuning data automatically and intelligently. To this end, We propose GENIXER, as illustrated in Fig. 3, 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.

### 3.1 Instruction Data Collection

In accordance with the prevalence and practical relevance of real-world multi-modal tasks, we carefully select nine representative multimodal tasks as listed in Tab. 1 for training GENIXER, 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,  $Q \rightarrow C^{Box}A$  and Referential Dialogue (RD). We divide these tasks into two categories, **general tasks** and **grounding tasks**, and display some examples in Fig. 4 for visual inspection.

### 3.2 Instruction Template Design

In an automatic data generation context, where image content is agnostic, preemptively determining the optimal data type for generation becomes particularly daunting, especially when it involves large-scale data production for model training purposes. For instance, in a landscape picture without any humans or objects, it is hard to ask a high-quality multi-choice question, and it is easy to generate an adversarial question such as “*Is there a boat in the image?*” Hence, it is imperative to delineate two key modes for a robust data-generative MLLM: 1) type-agnostic data generation and 2) type-specific data generation. **Two-level Instructions.** To achieve this goal, we propose a two-level instruction template for achieving controllable question-answer generation at the inference stage. Specifically, the complete instruction design is as follows:

<s> SYSTEM MESSAGE. USER: <image> **General 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 **General Instruction**, we meticulously compile a list of instructions, and for each piece of training data, we randomly choose one example to use during the training process. For instance, “*Please provide a clear and direct question and answer after examining the image*”. This instruction plays a role in allowing the model to generate diverse types of instruction tuning data without any specific condition referred to as mode 1.

Then, for **Specific Instruction**, it denotes designating the specific question type, such as “*This is a Common VQA task*”, allowing us to control the output type of the generated data referred to as mode 2.

**Controlling Constant.** During the training phase, we set a constant  $\tau$  for controlling the ratio of training samples that are exclusive with **General 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

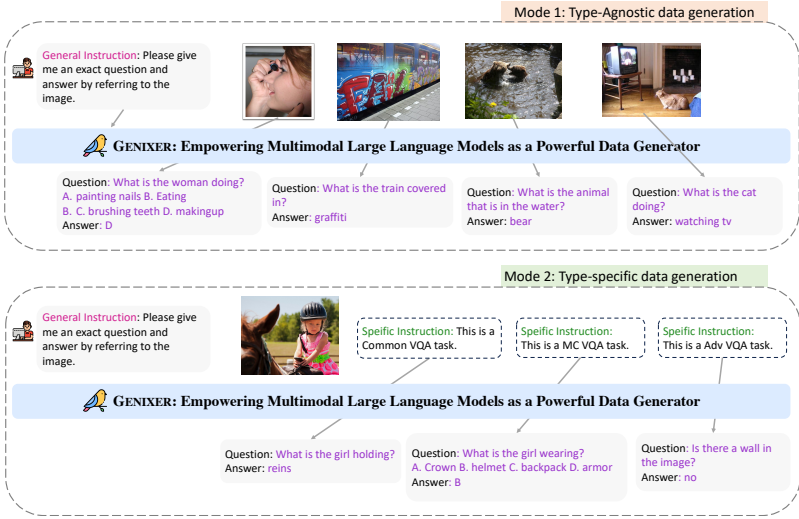


**Fig. 4:** Selected examples generated from GENIXER<sub>L</sub> and GENIXER<sub>S</sub>. The examples include Common VQA, Adv VQA, MC VQA, MD, and five grounding tasks.

Fig. 5, 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

A robust model with multimodal reasoning ability is the cornerstone for building an instruction tuning data generator. Therefore, we resort to delineating the current remarkable MLLM as the base model and then fine-tuning it with our developed two-level instruction template. Consequently, we present two data generators, GENIXER<sub>L</sub> and GENIXER<sub>S</sub>, to respectively generate instruction tuning data for the two data categories, general tasks, and grounding tasks as summarized in Tab. 1.



**Fig. 5:** A demonstration of two proposed instruction modes during the inference phase.

**Overall Framework of GENIXER<sub>L</sub> and GENIXER<sub>S</sub>.** For brevity, we denote the MLLM model as  $F_M$ , the input general and specific instructions as  $X_G$  and  $X_S$ , respectively. Then, given an image  $X_I$ , our target is to enforce  $F_M$  to generate question  $X_q$  and corresponding answer  $X_a$ :

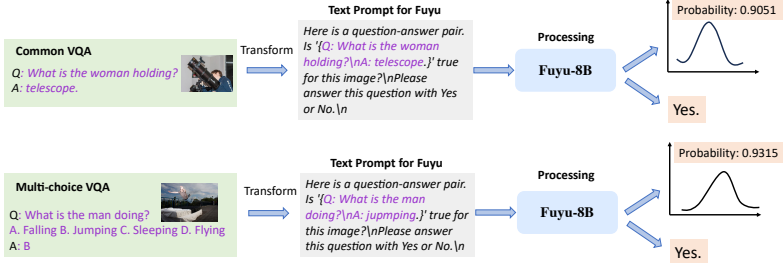
$$X_q, X_a = F_M(X_G, X_S, X_I). \quad (1)$$

To this end, we follow previous MLLMs [13, 30], and design the training objective in an autoregressive manner:

$$\max \sum_{i=1}^L \log p(X_o | (X_G, X_S, X_I)) = \prod_{i=1}^L p_\theta(x_i | (X_G, X_S, X_I, X_{o, < i})), \quad (2)$$

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.  $\theta$  denotes the total trainable parameters in  $F_M$  (e.g., the parameters of projector and LLM with LLaVA1.5 backbone). In the following, we select LLaVA1.5 and Shikra to generate two kinds of data because of their SoTA performance in the corresponding areas. It is noteworthy that GENIXER is flexible and applicable to work with any strong MLLM.

**Training of GENIXER<sub>L</sub>.** GENIXER<sub>L</sub> trains the pretrained LLaVA1.5 [30], a SoTA model for four kinds of general tasks, including Common VQA, Adv VQA, MC VQA, and MD. As summarized in Tab. 1, we only sample a subset of data for these four data types to improve training efficiency. The  $\tau$  set to 0.2, 0.2, 0.5, 0.2, respectively. The different ratios are because of the different data sizes of these four data types. To keep the training data balanced, we manually choose the values. Finally, we use AdamW [32] optimizer with a learning rate



**Fig. 6:** The illustration of proposed Fuyu-driven data filtering framework. The outputs of the framework compose a probability and a direct answer.

of  $1 \times 10^{-5}$  and a batch size of 128 to fine-tune the pretrained LLaVA1.5 for one epoch which takes about 14 hours.

**Training of GENIXER<sub>S</sub>.** Here, we aim to train a data generator for grounding tasks such as region-based VQA that involve the object coordinates by referring to the corresponding objects. To ensure the generation, we utilize the entirety of the RefCOCO [22] and VG [23] datasets. Nonetheless, as indicated in Tab. 1, the dataset size for RD is relatively small. To counteract potential biases in the generation, GENIXER<sub>S</sub> that finetunes a SoTA model Shikra [8] on region-based tasks adopts two-phased training. The first phase focuses on the REC and REG data generation. In the second phase, GENIXER<sub>S</sub> adds PointQA,  $Q \rightarrow C^{Box}A$ , and RD data while deliberately reducing REC and REG data for data trade-off. To train the GENIXER<sub>S</sub>, we set the  $\tau$  equal to 0.5 for each data type. We utilize the AdamW optimizer, applying learning rates of  $3 \times 10^{-5}$  for the first phase and  $1 \times 10^{-5}$  for the second phase. The batch sizes are set to 128 and 64, respectively.

### 3.4 Automatic Data Generation and Filtering

Here, we introduce how we use GENIXER<sub>L</sub> and GENIXER<sub>S</sub> to generate data for GENIXER<sub>L</sub> and GENIXER<sub>S</sub>, respectively. Additionally, recognizing that there are no ready-made state-of-the-art models available for the VQA data evaluation, we introduce two methods to assess the quality of the raw generated data. This process allows us to identify and select high-quality data for instruction tuning, which is essential for subsequent model training.

#### Data Generation and Filtering of GENIXER<sub>L</sub> On The General Tasks.

The diversity of image resources and effective data filtering are two key steps to obtaining high-quality instruction tuning data. Here, we first describe the image sources used for the data generation with GENIXER<sub>L</sub>. To prosper the diversity of current instruction tuning data, we utilize the mixed dataset comprising 558K images from the LAION [45], CC3M [46], and SBU [41], as described in [31], and further adopt the 830K images from the original SBU datasets. These different image resources can help to generate more diverse instruction tuning data. We directly feed these 1.4M images to the trained GENIXER<sub>L</sub> and then designate the specific instruction  $X_S = \text{“This is the Common VQA”}$

*task*”, and finally we produce the 1.4M raw VQA triplets pending for the next filtering.

To assess the quality of general task data, we design a Fuyu-driven data filtering framework to automatically and effectively filter the samples generated by GENIXER<sub>L</sub>. As illustrated in Fig. 6, 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

This framework is applicable to filter the four general tasks defined in Sec. 3.1. To assess the data quality of 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 [4], as shown in Fig. 6. For the MD type, we decompose multi-turn dialogue into individual single-turn VQA instances, processing these in the same manner as Common VQA.

To avoid the potential bias to output “Yes” of Fuyu-8B, instead of directly getting the words Yes or No as the final results, we additionally calculate the probability of predicting the “Yes”. To this end, we restrict the output to either Yes or No and calculate the probability via the output logits as follows:

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

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

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

where  $S^n$  is the sign of keeping or removing the current sample, and  $P(Y_r^n)$  denotes the probability of the result “Yes” of  $n$ -th visual-question-answer triplet. 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<sub>S</sub> on The Grounding Tasks.** Similarly, our GENIXER<sub>S</sub> adopts the same image resources in general tasks for generating grounding-based instruction tuning data. After feeding the 1.4M image corpus to GENIXER<sub>S</sub> 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 [20] 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**.

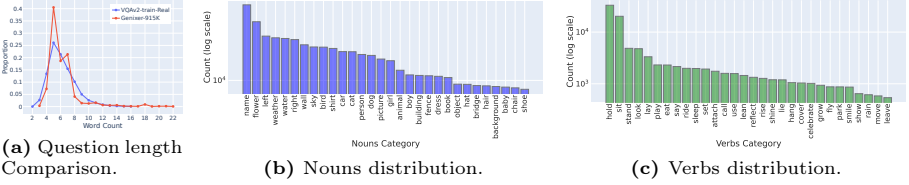
## 4 Experiments

In Sec. 3.4, we use two trained generators  $\text{GENIXER}_L$  and  $\text{GENIXER}_S$  to generate **Genixer-915K** and **Genixer-350K** for general and grounding tasks, respectively. Here, we evaluate the quality of these data from several aspects, including statistical analysis, human evaluation, evaluation via training MLLMs, ablation study, visualizations, and user study.

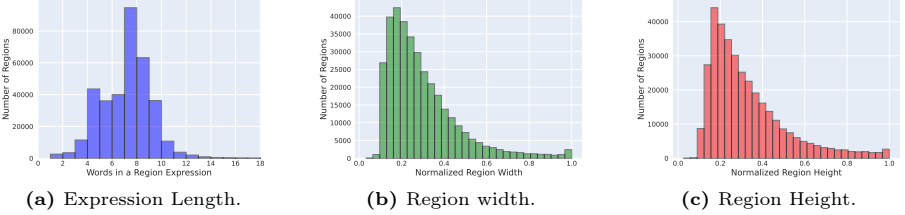
### 4.1 Statistical Analysis

To evaluate the generation quality of  $\text{GENIXER}_L$ , we conduct a comparative analysis using the VQAv2 [16] training set as a comparison. Fig. 7 (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. 7 (b) and (c), showcases the diverse vocabulary present in the generated questions. To assess data quality, we utilize Flickr30K [55] as the image source and employ  $\text{GENIXER}_L$  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. 9, affirming the high quality of data produced by  $\text{GENIXER}_L$  in generating diverse instruction tuning data.

The statistics of the generated dataset Genixer-350K by  $\text{GENIXER}_S$  are presented in Fig. 8, 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.



**Fig. 7:** The statistics of VQA-like dataset Genixer-915K.



**Fig. 8:** The statistics of REC-like dataset Genixer-350K.

## 4.2 Human Evaluation

We conduct the human evaluation to manually analyze the generated question type and corresponding correctness. We employ GENIXER<sub>L</sub> 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 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 held-in dataset is slightly higher than held-out dataset.

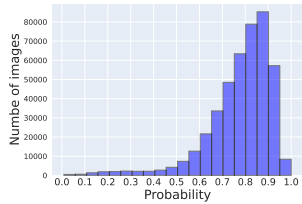
## 4.3 Evaluation via Training MLLMs

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

**Evaluation on General Tasks.** Considering the data size (665K) is small that is used for training LLaVA1.5 and the data balance during the finetuning stage, we proceed to add the Genixer-915K dataset into the pretraining phase, adhering to the same training protocols utilized by LLaVA1.5. Furthermore, we enriched the fine-tuning stage with an additional 8K VQA-like data points, selected based on Fuyu-8B’s probability range of 0.5 to 0.7. This probability range was chosen based on findings that higher probabilities correlate with simpler, more straightforward VQA instances. By selecting data within this

**Table 2:** Fuyu-8B evaluation result on Flickr30K image dataset. Accuracy refers to the “Yes” prediction. Prob. represents the probability.

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



**Fig. 9:** 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 [43]	31,783	275,775	–
RefCOCOg [36]	26,711	54,822	8.43
RefCOCO [22]	19,994	50,000	3.61
RefCOCO+ [22]	19,992	49,856	3.53
Visual Genome [23]	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

**Table 5:** Comparison with SoTA methods on 12 benchmarks. \* represents the train set used in training. All abbreviated names of benchmarks are following [30]. † indicates results we reproduced since the original results could not be replicated, as mentioned by the author in the corresponding GitHub issue.

Method	LLM	VQA <sup>v2</sup>	GQA	VizWiz	SQA <sup>1</sup>	VQA <sup>T</sup>	POPE	MME	MMB	MMB <sup>CN</sup>	SEED <sup>1</sup>	LLaVA <sup>W</sup>	MM-Vet
BLIP-2 [27]	Vicuna-13B	41.0	41.3	19.6	61.0	42.5	85.3	1293.8	–	–	49.7	38.1	22.4
InstructBLIP [13]	Vicuna-7B	–	49.2	34.5	60.5	50.1	–	–	36.0	23.7	58.8	60.9	26.2
InstructBLIP [13]	Vicuna-13B	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
Shikra [8]	Vicuna-13B	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
IDEFICS-9B [24]	LLaMA-7B	50.9	38.4	35.5	44.2	25.9	–	–	48.2	25.2	44.5	–	–
IDEFICS-80B [24]	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	<b>63.8</b>	–	–	38.2	7.4	62.3	–	–
Qwen-VL-Chat [3]	Qwen-7B	78.2*	57.5*	38.9	68.2	<u>61.5</u>	–	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	<b>65.4</b>	<b>31.1</b>
LLaVA-1.5+G-910K(ours)	Vicuna-7B	<b>79.1*</b>	<b>63.1*</b>	<b>53.8</b>	<b>69.7</b>	59.0	<b>87.3</b>	<b>1502.7</b>	<b>65.3</b>	<b>59.4</b>	<b>66.6</b>	<u>64.0</u>	<u>30.1</u>

specific range, our model is able to 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.

**Evaluation on Grounding Tasks.** Here, we adopt Shikra [8] as the model to evaluate the quality of Genixer-350K REC-like data produced by GENIXERs. Adhering to Shikra’s published code, we incorporate our synthetic data into

**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	
OFA-L [50]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
UNITER [10]	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	76.17
VILLA [14]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	77.77
UniTAB [53]	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	80.89
MDETR [21]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	81.81
Shikra [8]	87.01	90.61	80.24	81.60	87.36	72.12	<b>82.27</b>	82.19	82.92
Shikra <sub>+G-350K</sub> (ours)	<b>87.48</b>	<b>91.05</b>	<b>81.77</b>	<b>81.89</b>	<b>87.43</b>	<b>73.14</b>	81.99	<b>83.15</b>	<b>83.49</b>

**Table 7:** Performance of GENIXER<sub>L</sub> on 6 representative benchmarks.

Method	Setting	VQAv2	GQA	VizWiz	SQA <sup>I</sup>	POPE	SEED <sup>I</sup>
LLaVA1.5	-	78.5	62.0	50.0	66.8	85.9	66.2
GENIXER <sub>L</sub>	ZS	79.7 <sub>+1.2</sub>	61.3 <sub>-0.7</sub>	50.8 <sub>+0.8</sub>	65.6 <sub>-1.2</sub>	86.0 <sub>+0.1</sub>	65.6 <sub>-0.6</sub>
GENIXER <sub>L</sub>	MixT	80.2 <sub>+1.7</sub>	63.1 <sub>+1.1</sub>	54.1 <sub>+4.1</sub>	67.1 <sub>+0.3</sub>	87.5 <sub>+1.6</sub>	67.1 <sub>+0.9</sub>

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. 2.

**Performance on Genixer.** As illustrated in Tab. 2 and Fig. 9, GENIXER<sub>L</sub> showcases a superior capability of generating high-quality data. It is natural for us to investigate the performance of GENIXER<sub>L</sub>. Accordingly, we evaluate GENIXER<sub>L</sub> 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<sub>L</sub>, we proceeded to retrain the GENIXER<sub>L</sub> with the mixture of the 665K finetuning dataset used in LLaVA1.5 and the datasets for training GENIXER<sub>L</sub> 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.

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

Dataset	VQAv2	GQA	VizWiz	SQA <sup>†</sup>	POPE	SEED <sup>†</sup>
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
Genxier-915K	<b>79.1</b>	<b>63.1</b>	<b>53.8</b>	<b>69.7</b>	<b>87.3</b>	<b>66.6</b>

**Table 9:** The effect of different probability threshold  $\lambda$ .

Setting	Size	VQAv2	GQA	VizWiz	SQA <sup>†</sup>	POPE	SEED <sup>†</sup>
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	<b>79.1</b>	<b>63.1</b>	<b>53.8</b>	<b>69.7</b>	<b>87.3</b>	<b>66.6</b>

**Probability Range.** Tab. 9 investigates the impact of the probability threshold during data filtering in Section 3.4 on the data quality. By varying  $\lambda$ , we observe that higher values of  $\lambda$  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.

## 4.5 Visualizations

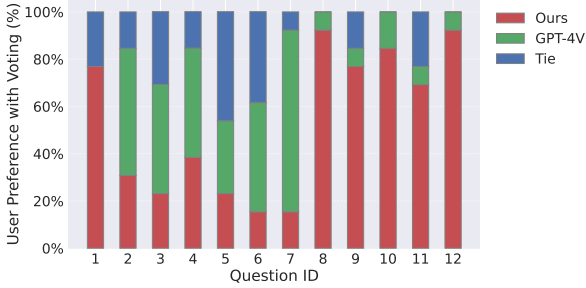
We sample some examples of general tasks such as Common VQA, MC VQA, and MD in Fig. 11, 12, 13, respectively. The examples of grounding tasks are shown in Fig. 14 and 15. These visualizations demonstrate that our GENIXER<sub>L</sub> and GENIXER<sub>S</sub> have the exceptional capability of generating diverse instruction tuning data. The generated VQA triplets in Fig. 11 show that GENIXER<sub>L</sub> is capable of generating diverse types of data such as color, Yes/No, and counting types. The generated samples of Fig. 12 show the ability to generate multi-choice VQA data at a high-quality level. As for generating MD data, we display four examples in Fig. 13. The results demonstrate the ability to generate long semantic sentences when answering the questions. Furthermore, we also compare the REC data generation ability with GPT-4V [40] in Fig. 16. It indicates that GPT-4V is suboptimal for generating grounding tasks such as REC.

## 4.6 User Study

We conducted a user preference study to evaluate the generation quality between GENIXER and GPT-4V. For this study, we selected 12 samples, comprising 4 Common VQA samples, 3 MC VQA samples, and 5 REC samples. Fig. 10 summarizes the statistical analysis of 13 valid surveys. (1) The first seven columns (excluding the first one) reveal that GPT-4V was the primary preference among users, while between 20% to 40% of users preferred the data generated by GENIXER. Notably, a significant number of users selected “Tie” indicating that the data quality we generated is comparable to that of GPT-4V. (2) In analyzing the last five columns, our generated REC samples were more favored than those of GPT-4V, as the evidence shown in Fig. 16.

## 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 instruc-



**Fig. 10:** User preference with voting for comparing the generated data quality between GENIXER and GPT-4V.

tion tuning data by leveraging current MLLMs. We instantiate GENIXER into two data-generative MLLMs,  $\text{GENIXER}_L$  and  $\text{GENIXER}_S$ , 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 model, such as 13B or 34B. But 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 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.

## 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, 4, 5, 13



## Common VQA



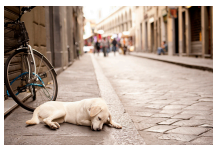
**Question:** What is the man doing?  
**Answer:** skateboarding



**Question:** What is the name of the street on the bottom sign?  
**Answer:** ross



**Question:** How many clocks are there? **Answer:** 1



**Question:** What is the dog sleeping on? **Answer:** sidewalk



**Question:** What is the man doing?  
**Answer:** snowboarding



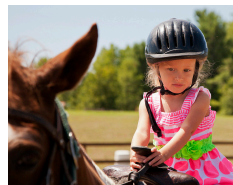
**Question:** How many people are in the water? **Answer:** 5



**Question:** What is the woman holding? **Answer:** cell phone



**Question:** What is the color of the fridge? **Answer:** silver



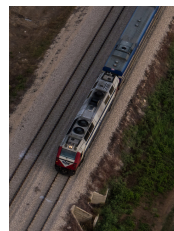
**Question:** What is the girl wearing? **Answer:** helmet



**Question:** What is the graffiti on? **Answer:** train




**Question:** What is the man holding? **Answer:** tie




**Question:** Is the train in motion? **Answer:** yes

**Fig. 11:** Some examples of Common VQA generated by GENIXER<sub>L</sub>.


Multi-choice VQA




**Question:** What is the bear sitting on?  
A. car seat B. car floor C. car door D. car console  
**Answer:** B




**Question:** What is the tablecloth made of?  
A. Paper B. plastic C. nylon D. Lace  
**Answer:** D




**Question:** What is the orange vegetable called?  
A. carrot B. pumpkin C. squash D. sweet potato  
**Answer:** A




**Question:** What is the red object used for?  
A. fighting fires B. drinking C. eating D. Bathing  
**Answer:** A




**Question:** What is the man holding?  
A. Egg B. umbrella C. kitten D. apple  
**Answer:** B




**Question:** What is the name of the building on the right?  
A. Grand B. thorne-sec C. made frames D. great Britain  
**Answer:** A



**Question:** What is the man doing?  
A. Jumping B. falling C. balancing D. rolling  
**Answer:** C



**Question:** What is the name of the river?  
A. thames B. amazon C. nile D. ganges  
**Answer:** A



**Question:** What is the man in black's position?  
A. umpire B. catcher C. pitcher D. Referee  
**Answer:** A

Fig. 12: Some examples of Multi-choice VQA generated by GENIXER<sub>L</sub>.

### Multi-turn Dialogue

**User:** Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.



**Question:** What is the man in the image doing?

**Answer:** The man in the image is talking on his cell phone while standing on a crowded sidewalk.

**Question:** What is the man wearing on his head?

**Answer:** The man is wearing a baseball cap on his head.

**Question:** What is the man's appearance?

**Answer:** The man has a beard and appears to be older.

**Question:** What is the setting of the image?

**Answer:** The setting of the image is a crowded city street, with people walking around and engaging in various activities.

**Question:** What can be inferred about the man's conversation on the phone?

**Answer:** While it is impossible to know the exact content of the man's conversation, we can infer that he is likely discussing something important or serious, as he is talking on the phone in a crowded and noisy environment.



**Question:** What is the boy doing in the image?

**Answer:** The boy in the image is throwing a green frisbee in a park.

**Question:** What is the color of the frisbee the boy is throwing?

**Answer:** The color of the frisbee the boy is throwing is green.

**Question:** What is the setting of the image?

**Answer:** The setting of the image is a park, with the boy standing in the grass.

**Question:** What is the boy wearing?

**Answer:** The boy is wearing a grey sweater.

**Question:** What is the age range of the boy in the image?

**Answer:** The boy in the image is a young child, but the exact age cannot be determined from the image alone.



**Question:** What are the main appliances visible in the kitchen?

**Answer:** The main appliances visible in the kitchen are a sink, a stove, and a microwave.

**Question:** What is the color of the kitchen cabinets?

**Answer:** The kitchen cabinets are brown.

**Question:** What is the condition of the kitchen?

**Answer:** The kitchen is clean and ready for use.

**Question:** What is the condition of the kitchen counter?

**Answer:** The kitchen counter is cluttered with various items.

**Question:** Is there a window in the kitchen?

**Answer:** Yes, there is a window in the kitchen.



**Question:** What is the young boy doing in the image?

**Answer:** The young boy in the image is flying a kite in a field.

**Question:** What is the color of the boy's shirt?

**Answer:** The boy is wearing a blue shirt.

**Question:** What is the condition of the field where the boy is flying the kite?

**Answer:** The field where the boy is flying the kite is covered with dry grass.

**Question:** What is the boy's expression while flying the kite?

**Answer:** The boy is smiling while flying the kite, indicating that he is enjoying the activity.

**Question:** What can you infer about the boy's experience with flying a kite?

**Answer:** Based on the image, it can be inferred that the young boy is likely a beginner or has some experience with flying a kite, as he is actively engaged in the activity and enjoying the process.

**Fig. 13:** Some examples of Multi-turn Dialogue generated by GENIXER<sub>L</sub>.

### Referring Expression Comprehension



**User:** I'd appreciate a precise question and your answer. Keeping the image in mind.  
 \nThis is a Referring Expression Comprehension (REC) task.

**Question:** I'd like to request the coordinates of a man in a blue shirt and gray pants that is bent over with his hands on his knees within the photo.

**Answer:** [0.453,0.320,0.784,0.727]



**User:** I'd appreciate a precise question and your answer. Keeping the image in mind.  
 \nThis is a Referring Expression Comprehension (REC) task.

**Question:** I'd like to request the coordinates of red and white awning within the photo.

**Answer:** [0.000,0.401,0.382,0.602]

### Referring Expression Generation



**User:** I'd appreciate a precise question and your answer. Keeping the image in mind.  
 \nThis is a Referring Expression Generation (REG) task.

**Question:** What are the unique aspects of the selected rectangular area [0.668,0.514,0.988,0.919] in image?

**Answer:** A group of people crossing the street.

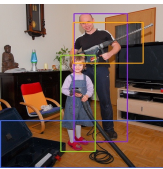


**User:** I want to know the question and answer. Refer to the image.  
 \nThis is a Referring Expression Generation (REG) task.

**Question:** What are the unique aspects of the selected rectangular area [0.494,0.372,0.648,0.634] in image?

**Answer:** pink cup being held by child.

### $Q \rightarrow C^{\text{box}} A$



**User:** Can you give me a straightforward question and answer? Include questions asking about the visual content of the image.  
 \nThis is a QCA with box task.

**Question:** Is there an event "a man and a girl using a vacuum cleaner together" in the picture? **Answer:** In the picture, a man [0.447,0.058,0.767,0.838] and a little girl [0.367,0.312,0.573,0.898] are both holding a vacuum cleaner [0.441,0.112,0.862,0.360]. They are standing together and working together on cleaning the floor [0.003,0.710,1.000,0.998]. Based on... Therefore, the answer is yes.

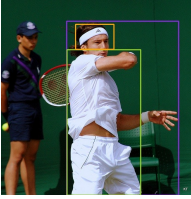


**User:** Can you provide a clear and direct question and answer by analyzing the image?  
 \nThis is a QCA with box task.

**Question:** Is there an event "two men wearing swimsuits playing with a frisbee on the beach" in the picture? **Answer:** The image shows two men [0.228,0.334,0.552,0.544 0.594,0.396,0.878,0.826] on the beach [0.002,0.510,1.000,1.000]. However, there is no evidence of them wearing swimsuits or playing with a frisbee. The men are engaged in a kung fu moves while wearing swimsuits. Therefore, the answer is no.

**Fig. 14:** Some examples of grounding tasks generated by GENIXERS.

### Referential Dialogue



**User:** Can you give me a straightforward question and answer? Include questions asking about the visual content of the image.\nThis is a Referential Dialogue task.

**Question:** Is the man [0.343,0.108,0.931,1.000] wearing any visible accessories or hats?

**Answer:** The man [0.343,0.108,0.931,1.000] is only wearing white clothes [0.346,0.254,0.731,1.000] and a white headband [0.391,0.124,0.591,0.252]. There are no visible accessories or hats in the image.



**User:** I require a brief and clear question and answer regarding the image.\nThis is a Referential Dialogue task.

**Question:** Is the girl [0.182,0.002,1.000,1.000] holding the bucket [0.218,0.686,0.532,0.942]?

**Answer:** The girl [0.182,0.002,1.000,1.000] is not holding the bucket [0.218,0.686,0.532,0.942]; it is a green plastic shovel [0.214,0.690,0.583,0.916] that she is holding in her hand [0.184,0.704,0.375,0.822].

### PointQA



**User:** Can you provide a clear and direct question and answer by analyzing the image?\nThis is a PointQA task.

**Question:** Which is the yellow soccer cleat of the player?

**Candidates:** [0.200,0.434,0.393,0.598] [0.171,0.744,0.584,0.936] [0.242,0.210,0.490,0.864] [0.450,0.820,0.549,0.944]

**Answer:** [0.450,0.820,0.549,0.944]



**User:** Can you provide a clear and direct question and answer by analyzing the image? \nThis is a PointQA task.

**Question:** What color is this shirt? [0.112,0.249,0.330,0.599]

**Answer:** blue



**User:** Can you provide a clear and direct question and answer by analyzing the image? \nThis is a PointQA task.

**Question:** What color is this helmet?[0.412,0.231,0.556,0.402] **Answer:** yellow



**User:**Can you provide a clear and direct question and answer? \nThis is a PointQA task.

**Question:** What color is this shirt? [0.889,0.490]

**Answer:** white



**User:** Can you provide a clear and direct question and answer? \nThis is a PointQA task.

**Question:** What color is this jacket? [0.492,0.536]

**Answer:** red

Fig. 15: Some examples of grounding tasks generated by GENIXER<sub>S</sub>.





**Fig. 16:** Comparison of generating REC-like data between GENIXER<sub>S</sub> and GPT-4v [40].



4. Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., Taşlılar, S.: Introducing our multimodal models (2023), <https://www.adept.ai/blog/fuyu-8b> **3**, 10
5. 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**
6. 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) **4**
7. 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) **4**
8. 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**, **9**, **13**, **14**
9. 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**
10. 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) **14**
11. 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/> **4**
12. 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) **4**
13. 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 (2023) **1**, **2**, **4**, **8**, **13**
14. 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) **14**
15. 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**, **4**
16. 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) **2**, **4**, **11**
17. 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) **4**
18. 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**, **4**
19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: *CVPR* (2019) **4**

20. 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> 3, 11
21. 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) 14
22. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 9, 13
23. 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) 9, 13
24. 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) 13
25. 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
26. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023) 5
27. 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) 4, 13
28. 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
29. 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
30. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 2, 3, 5, 8, 13
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 1, 4, 9
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=Bkg6RiCqY7> 8
33. 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) 4
34. Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models (2023) 4
35. Mani, A., Yoo, N., Hinthorn, W., Russakovsky, O.: Point and ask: Incorporating pointing into visual question answering. arXiv preprint arXiv:2011.13681 (2020) 5
36. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 13
37. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019) 2, 4

38. 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) [4](#)
39. OpenAI: Gpt-4 technical report (2023) [1](#), [5](#)
40. OpenAI: Gpt-4v(ision) system card (2023), [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf) [2](#), [3](#), [5](#), [15](#), [22](#)
41. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. NeurIPS (2011) [9](#)
42. 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](#), [4](#), [5](#)
43. 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) [13](#)
44. 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](#)
45. 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) [9](#)
46. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) [9](#)
47. 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](#)
48. 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) [4](#)
49. 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) [4](#)
50. 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) [14](#)
51. 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 (2023) [4](#), [5](#)
52. 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) [4](#), [5](#)
53. 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) [14](#)
54. 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](#), [2](#), [4](#)
55. 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](#), [11](#)

- 56. 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](#)
- 57. 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) [4](#)
- 58. 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](#), [4](#)