

# InstructionGPT-4: A 200-Instruction Paradigm for Fine-Tuning MiniGPT-4

Lai Wei<sup>1</sup>   Zihao Jiang<sup>1</sup>   Weiran Huang<sup>1,\*</sup>   Lichao Sun<sup>2</sup>

<sup>1</sup>Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University

<sup>2</sup>Lehigh University

## Abstract

Multimodal large language models acquire their instruction-following capabilities through a two-stage training process: pre-training on image-text pairs and fine-tuning on supervised vision-language instruction data. Recent studies have shown that large language models can achieve satisfactory results even with a limited amount of high-quality instruction-following data. In this paper, we introduce InstructionGPT-4, which is fine-tuned on a small dataset comprising only 200 examples, amounting to approximately 6% of the instruction-following data used in the alignment dataset for MiniGPT-4 [1]. We first propose several metrics to access the quality of multimodal instruction data. Based on these metrics, we present a simple and effective data selector to automatically identify and filter low-quality vision-language data. By employing this method, InstructionGPT-4 outperforms the original MiniGPT-4 on various evaluations (e.g., visual question answering, GPT-4 preference). Overall, our findings demonstrate that less but high-quality instruction tuning data is efficient to enable multimodal large language models to generate better output.

## 1 Introduction

GPT-4 [2] has showcased its powerful prowess in generating highly detailed and precise descriptions of images, signaling a new era of language and visual processing. Thus, GPT-4 like Multimodal Large Language Models (MLLMs) have recently emerged as a prominent research area, harnessing powerful Large Language Models (LLMs) as a cognitive framework for conducting multimodal tasks. The remarkable and unexpected capabilities exhibited by MLLMs surpass those of traditional methods, indicating a potential pathway towards artificial general intelligence. To achieve this, massive image-text pairs and vision-language tuning data have been employed to train connectors (e.g., MiniGPT-4 [1], LLaVA [3], LLaMA-Adapter V2 [4]) between frozen LLMs (e.g., LLaMA [5] and Vicuna [6]) and visual representations (e.g., CLIP [7] and BLIP-2 [8]).

MLLMs are usually trained in two stages: pre-training and fine-tuning. Pre-training helps MLLMs gain a large amount of knowledge, while fine-tuning teaches models to better understand human intentions and generate accurate responses. Recently, instruction tuning on large-scale datasets has served as a powerful fine-tuning technique to empower MLLMs with enhanced vision-language understanding and instruction-following abilities [9–11]. It facilitates the alignment of models with human preferences, enabling the generation of desired outputs in response to various instructions. A constructive direction to develop instruction tuning involves the introduction of image caption, visual question answering (VQA), and visual reasoning datasets

---

\*Correspondence to Weiran Huang (weiran.huang@outlook.com).

in the fine-tuning stage. Previous works, including InstructBLIP [12] and Otter [13], have shown promising results by leveraging a collection of vision-language datasets for visual instruction tuning.

However, it has been observed that commonly used multimodal instruction-tuning datasets surprisingly contain numerous low-quality instances with incorrect or irrelevant responses. Such data can mislead and negatively impact the performance of the model. This issue has prompted researchers to delve into the possibility of achieving robust performance using a small quantity of high-quality instruction-following data. Encouragingly, recent studies have substantiated the promising potential of this approach. Zhou et al. [14] introduce LIMA, a language model fine-tuned with carefully curated high-quality data, selected by human experts. This study has shown that large language models can achieve satisfactory results even with a limited amount of high-quality instruction-following data. Nevertheless, the process of identifying appropriate high-quality datasets for fine-tuning multimodal language models lacks clear guidelines.

Building upon these foundations, we propose a robust and effective data selector that automatically identifies and filters low-quality vision-language data, ensuring that our model is trained on the most relevant and informative samples. The key focus of our study lies in exploring the efficacy of reduced but high-quality instruction-tuning data in fine-tuning multimodal large language models. Additionally, our work introduces several novel metrics tailored for assessing the quality of multimodal instruction data. Data selector computes a weighted score combining CLIP Score [7], GPT Score [15], Reward Score [16] and Answer Length of each vision-language data after conducting spectral clustering on images. By applying this selector to the original 3.4K data used to fine-tune MiniGPT-4, we find that a majority of the data suffer from low-quality issues. Using the data selector, InstructionGPT-4 is fine-tuned on a much smaller but carefully selected subset of 200 data, which is 6% of the original scale, following the same training configuration of MiniGPT-4. This discovery is inspiring, as it demonstrates that the data quality in vision-language instruction tuning can outweigh the quantity. In addition, this shift towards prioritizing data quality presents a new and more efficient paradigm that can generally improve the fine-tuning of MLLMs.

In the subsequent sections, we provide an in-depth account of the experimental setup. Our evaluation of the fine-tuned MLLMs focuses on seven various and complex open-domain multimodal datasets, including Flickr-30k [17], ScienceQA [18], VSR [19], etc. Through rigorous experimentation, we demonstrate the superiority of InstructionGPT-4 across these diverse multimodal tasks by comparing the inference performance among models fine-tuned using data selector, randomly sampled dataset and whole dataset. Besides, we also employ GPT-4 as our judge for the evaluation. Specifically, we apply a prompt to turn GPT-4 into a judge comparing the responses of InstructionGPT-4 and the original MiniGPT-4 using the test set in LLaVA-Bench [3]. Despite being fine-tuned on a mere 6% of the original instruction-following data utilized in MiniGPT-4 [1], InstructionGPT-4 produces equal or preferable responses in 73% of the cases.

Our contributions are summarized as follows:

- In this paper, we are the first to show that less instruction data for better alignment is suitable for multimodal large language model by selecting 200 (nearly 6%) instruction-following data with high quality to train InstructionGPT-4.
- We propose a data selector that utilizes a simple and explainable principle to select high-quality multimodal instruction-following data for fine-tuning. This approach strives for both effectiveness and portability in the evaluation and curation of data subsets.
- We demonstrate this simple technique to work well in different tasks. Comprehensive results show that InstructionGPT-4 fine-tuned on 6% filtered data performs better than the original MiniGPT-4 in various tasks.

## 2 Related Works

**Visual Instruction Tuning.** Instruction tuning is a learning paradigm that fine-tunes pre-trained LLMs on datasets described by natural language instructions. Through this training method, the zero-shot abilities of LLMs can be significantly enhanced. The effectiveness of instruction tuning has been demonstrated by a set of research, including InstructGPT [20], ChatGPT and FLAN [21]. Inspired by this, several recent works aim at enabling LLMs to handle multimodal tasks with visual instruction tuning, such as MiniGPT-4 [1], LLaVA [3], LLaMA-Adapter [22] and InstructBLIP [12]. These works choose linear projection layers as the bridges between image encoders and LLMs, and perform visual instruction tuning either on self-instruct datasets [1, 3] or on existing multimodal datasets [12, 22].

**Instruction Curation.** To improve model performance after instruction tuning, some relevant works manage to filter low-quality instruction data or construct carefully curated examples during the fine-tuning stage, thereby enhancing model capabilities. Polite Flamingo [23] is trained to reconstruct high-quality responses from their automatically distorted counterparts and is subsequently applied to a vast array of vision-language datasets for response rewriting. LIMA [14] shows that fine-tuning a strong pre-trained language model on a few curated and high-quality examples can produce remarkable, competitive results on a wide range of prompts. Thus, several recent works [15, 24] have developed instruction quality evaluation methods for measuring the quality of instruction datasets, such as using reward models, computing the length of instruction, and acquiring ChatGPT for rating, to filter low-quality data for alignment. In contrast to recent works such as LIMA [14], which relies on human annotation, or Polite Flamingo [23], which needs to rewrite responses, our work aims to present a multimodal data quality evaluation principle for selecting proper data from raw dataset used during fine-tuning.

## 3 Methodology

We aim to propose a simple and migratable data selector to automatically curate a subset from the original fine-tuning dataset. Hence, we define a selecting principle that focuses on the diversity and quality of the multimodal dataset and streamline the selection process as follows.

### 3.1 Selecting Principle

Selecting useful multimodal instruction data is crucial for effectively training MLLMs. We propose two key principles for selecting optimal instruction data: diversity and quality. To achieve diversity, we use a clustering mechanism on image embeddings to categorize data into distinct groups. For assessing quality, we adopt several key indicators to efficiently evaluate multimodal data.

**Diversity.** As most of the knowledge is obtained during the pre-training stage for MLLMs, it is necessary to gain better alignment abilities by training on diverse vision-language instruction data. We adopt spectral clustering on the image embeddings encoded by CLIP [7] to divide the data into ten categories. Spectral clustering is a popular technique used in image analysis and computer vision to understand and analyze the diversity of images within a dataset. It is an unsupervised learning method that aims to group similar images together based on their visual characteristics by constructing a similarity matrix, projecting features onto a lower-dimensional space, and applying  $K$ -means clustering. Our ablation study is detailed in Section 5.3.

**Quality.** Vision-language instruction data teaches the multimodal model to follow a certain pattern when interacting with users. Hence, the quality of these instruction-following data could be viewed as its ability to efficiently steer multimodal language models in learning to generate responses in a particular manner. Recently, Cao et al. [24] propose a linear quality rule and a bag of indicators for evaluating instruction-following data quality. We further propose our multimodal instruction selecting principle as follows. The relevant indicators for quantitatively evaluating data quality are shown in Table 1.

Indicators	Explanation
CLIP Score	The cosine similarity between image embedding and response text embedding. The CLIP Score serves as a measure of the alignment between the provided image and its accompanying caption. This score quantifies how well the caption accurately describes the visual content, ensuring that the image and text are in concordance.
Answer Length	The length of every answer in the multimodal dataset. The length metric gauges the extent of information encapsulated within the caption. A balanced and informative answer length is crucial to convey the desired instruction without being excessively verbose or overly concise.
Reward Score	Score from a reward model [16] that judges the human likeness to a response. The reward model is trained from human feedback to predict which generated answer is better judged by a human, given a question.
GPT Score	Score from ChatGPT to evaluate the quality of response. The GPT Score reflects the LLM’s assessment of the caption’s quality. This score is indicative of how effectively the generated caption adheres to the model’s language proficiency, considering factors such as grammar, semantics, and fluency.

Table 1: Quantitative indicators and explanations for evaluating instruction-following data quality. CLIP Score measures the suitability between the image and caption. Answer Length, Reward Score, and GPT Score measure the comprehensive quality of the caption.

### 3.2 Data Selector

Given a vision-language instruction dataset  $D$  of triplets  $x = (\text{image}, \text{instruction}, \text{answer})$  with  $x \in D$  and a pre-trained MLLM (e.g., MiniGPT-4 and LLaVA), our ultimate objective is to identify a subset  $S \subset D$  that, when utilized for fine-tuning, leads to the improvement of the pre-trained MLLM.

In order to select  $S$  from  $D$  and ensure its diversity, we first use a clustering algorithm (e.g., spectral clustering and  $K$ -means++) to separate the images in  $D$  into  $K$  categories. Suppose that the total amount of  $D$  is  $|D|$  and the  $i$ -th cluster’s amount is  $|D_i|$ . We set  $|S| = \alpha$  as the size of the target subset.

To assure the quality of the selected multimodal instruction data, we formulate a set of indicators for assessment in Table 1. For the triplets  $x$  in each cluster, we employ CLIP Score [7]  $C(x)$  from the pre-trained CLIP model to measure the matching degree between the image and answer. We also take the length of responses into consideration, which is  $L(x)$ , as longer output can contain more information. Besides, we apply Reward Score [16]  $R(x)$  from a reward model trained to predict which generated answer is better judged by a human when given a question. We prompt a powerful LLM (e.g., GPT-3.5-turbo and GPT-4 [2]) as an auto-grader rating each sample  $x \in D$  with a GPT Score  $G(x, p_G)$  wherein  $p_G$  is the rating instruction. We design  $p_G$  based on the GPT prompt from Alpargus [15], which is shown in Appendix B.

The final score  $F(x)$  for each triplet  $x$  can be formulated as:

$$F(x) = \lambda_1 C(x) + \lambda_2 L(x) + \lambda_3 R(x) + \lambda_4 G(x, p_G),$$

where each part of the score is rated out of 100, and  $\lambda_i, i \in \{1, 2, 3, 4\}$  refers to the weight (Table 2) of each score set manually.

---

**Algorithm 1** DATA SELECTOR

---

**Require:** Dataset  $D$ , weights  $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ , number of clusters  $K$ , subset size factor  $\alpha$

- 1: Compute image clusters  $D_1, D_2, \dots, D_K$  using a clustering algorithm on images in  $D$
- 2: **for**  $i = 1$  to  $K$  **do**
- 3:   **for**  $x$  in  $D_i$  **do**
- 4:     Compute CLIP Score  $C(x)$ , Answer Length  $L(x)$ , Reward Score  $R(x)$ , GPT Score  $G(x, p_G)$
- 5:     Set the weight  $\lambda_i, i \in \{1, 2, 3, 4\}$  for each score
- 6:     Compute final score  $F(x) = \lambda_1 C(x) + \lambda_2 L(x) + \lambda_3 R(x) + \lambda_4 G(x, p_G)$
- 7:   **end for**
- 8:   Compute  $|S_i| = \frac{\alpha \cdot |D_i|}{|D|}$
- 9:   Select top  $|S_i|$  samples from  $D_i$  based on  $F(x)$  to form  $S_i$
- 10: **end for**
- 11: Combine  $S_1, S_2, \dots, S_K$  to form  $S$
- 12: **return**  $S$

---

We sort  $x$  according to  $F(x)$  and select  $S_i$  from each cluster  $D_i$ . Each  $S_i$  contains top  $|S_i|$  triplets  $x$  based on  $F(x)$  from  $D_i$ , i.e.,

$$|S_i| = \frac{\alpha \cdot |D_i|}{|D|}, \quad S_i = \underset{V \subset D_i, |V|=|S_i|}{\operatorname{argmax}} \sum_{x \in V} F(x).$$

At last, we combine these  $K$  subsets:

$$S = S_1 \cup S_2 \cup \dots \cup S_K,$$

where  $S$  is the final dataset selected by the data selector. Algorithm 1 illustrates the overall pipeline.

## 4 Experimental Setup

### 4.1 Datasets

To prove the efficacy of our data selector, we also select  $\alpha$  vision-language instruction data randomly from the raw fine-tuning dataset [1] that contains 3439 data for alignment.

A comparative analysis is conducted among models that undergo fine-tuning using three distinct datasets: the unprocessed raw dataset, the randomly sampled subset, and the data selector-crafted dataset. For the ablation study, we first remove the clustering mechanism from the data selector and conduct selection directly based on the final scores of each triplet. This removal allows us to isolate and scrutinize the specific impact of the clustering component on the overall performance.

Variable	Category	Weight
$\lambda_1$	CLIP Score	0.53
$\lambda_2$	Answer Length	0.10
$\lambda_3$	Reward Score	0.10
$\lambda_4$	GPT Score	0.27

Table 2: Weight of each separate score for computing the final score of each triplet.

Furthermore, we evaluate models that are fine-tuned using subsets selected based on individual scoring metrics. Specifically, we assess models tuned with subsets chosen according to CLIP Score, Answer Length, Reward Score, and GPT Score, each in separation.

### 4.2 Fine-Tuning Settings

The vision-language dataset is categorized into  $K = 10$  clusters for data selection. We set the weights of each separate score manually in Table 2 and select diverse and high-quality multimodal data based on the final score

$F(x)$ . The final subset  $S$  from the data selector contains  $\alpha = 200$  vision-language instruction data, which is 6% of the original amount. We conduct all instruction tuning on pre-trained 7B MiniGPT-4 [1] and use the same fine-tuning hyperparameters as the original MiniGPT-4. Each fine-tuned model is evaluated on the evaluation dataset mentioned in Section 5.1.

Dataset Name	Description	Size
Flickr-30k	The Flickr30k dataset consists of 31K images collected from Flickr, each image has five ground truth captions.	1K (test)
NoCaps	The NoCaps dataset contains 15100 images with 166100 human-written captions for novel object image captioning.	4500 (val)
ScienceQA	ScienceQA is a multimodal benchmark containing multiple choice questions with a diverse set of science topics.	2017 (test)
OKVQA	OKVQA is a dataset about outside knowledge visual question answering. It contains 14055 open-ended question-answer pairs in total.	5046 (val)
VSR	Visual Spatial Reasoning (VSR) dataset contains a collection of caption-image pairs with true/false labels. We treat it as a VQA dataset by asking the model to answer True or False.	10972 (all)
VCR-OC	We construct a dataset using VCR validation images to evaluate fine-grained visual understanding by asking the model to count and attend to individual objects, which is decoupled from high-level semantics.	10000 (val)
VCR-MCI	We construct a dataset using VCR validation images to evaluate fine-grained visual understanding by asking the model to attend to individual objects and determine their existence, which is decoupled from high-level semantics.	10000 (val)

Table 3: Description of datasets used in our evaluation. We use CIDEr score for image captioning datasets, top-1 accuracy for VQA and VSR datasets, and accuracy metric for VCR datasets.

### 4.3 Evaluation

MLLMs are capable of capturing a wide range of multimodal patterns and relationships. Most are evaluated on publicly available datasets or judged by GPT-4 [2].

LVLm-eHub [25] is a comprehensive evaluation benchmark for publicly available MLLMs. Based on this platform, we evaluate the image captioning, visual spatial reasoning, visual commonsense reasoning, knowledge-grounded image description and visual question answering capabilities of MLLMs by investigating their zero-shot performance on various tasks. We choose Flickr-30k [17], NoCaps [26], ScienceQA [18], OKVQA [27], VSR [19], VCR Object Counting and VCR Multi-Class Identification [28] to evaluate the MLLMs’ zero-shot ability to generalize to new tasks without training the model, which is competent for large-scale evaluation. Table 3 provides an overview of the evaluation datasets.



LLaVA-Bench [3] collects a diverse set of 24 images with 60 questions in total, including indoor and outdoor scenes, memes, paintings, sketches, etc. It associates each image with a highly-detailed and manually-curated description and a proper selection of questions. We choose GPT-4 as a judge to compare the responses from MiniGPT-4 and InstructionGPT-4 given the images and instructions from LLaVA-Bench. The questions are categorized into three categories: conversation (simple QA), detailed description, and complex reasoning. The score is measured by comparing MLLM’s output against a reference answer. Such a design assesses the model’s robustness to different prompts.

## 5 Experimental Results

### 5.1 Benchmark Scores

Table 4 shows the performance comparison among MiniGPT-4 baseline model, MiniGPT-4 tuned from randomly sampled data, and InstructionGPT-4 with the data selector respectively. We observe that InstructionGPT-4 provides the ceiling performance on average. Specifically, InstructionGPT-4 demonstrates 2.12% improvement over the baseline model’s performance on ScienceQA, 2.49% on OKVQA and 4.19% on VCR-OC. Moreover, InstructionGPT-4 outperforms the model trained from random samples on all other tasks except VSR. By evaluating and contrasting these models in a range of tasks, we aim to discern their respective capabilities and ascertain the efficacy of our proposed data selector that can effectively identify high-quality data. This comprehensive analysis sheds light on the benefits of informed data selection in enhancing zero-shot performance across diverse tasks.

Datasets	MiniGPT-4 (3439 samples)	Random Selection (200 samples)	InstructionGPT-4 (200 samples)
Flickr-30k	<b>21.57</b> ( $\pm 0.13$ )	17.88 ( $\pm 0.26$ )	20.31 ( $\pm 0.07$ )
NoCaps	<b>42.43</b> ( $\pm 0.26$ )	38.61 ( $\pm 0.48$ )	41.63 ( $\pm 0.38$ )
ScienceQA	22.44 ( $\pm 0.29$ )	21.73 ( $\pm 0.45$ )	<b>24.56</b> ( $\pm 0.39$ )
OKVQA	38.20 ( $\pm 0.89$ )	36.60 ( $\pm 0.17$ )	<b>40.69</b> ( $\pm 0.51$ )
VSR	<b>41.20</b> ( $\pm 0.38$ )	40.72 ( $\pm 0.72$ )	39.85 ( $\pm 0.31$ )
VCR-OC	34.22 ( $\pm 0.11$ )	29.67 ( $\pm 1.01$ )	<b>38.41</b> ( $\pm 0.28$ )
VCR-MCI	<b>55.59</b> ( $\pm 0.21$ )	47.17 ( $\pm 0.98$ )	51.95 ( $\pm 0.40$ )
Average Score	36.52	33.20	<b>36.77</b>

Table 4: Comparison of zero-shot performance on different tasks. The Baseline refers to the original MiniGPT-4, while the Random Sample pertains to the model fine-tuned using randomly selected 200 data and InstructionGPT-4 denotes the the model fine-tuned using 200 data from the data selector. This assessment offers a valuable perspective on the efficacy of data selector in enhancing zero-shot performance across a range of tasks.

### 5.2 GPT-4 Evaluation

Given the presence of inherent position bias within LLMs as evaluators, wherein certain positions are favored over others [29], we have undertaken measures to address this concern. To mitigate such bias, we have conducted evaluations using both response orders — placing InstructionGPT-4’s generated response before and after MiniGPT-4’s response. To establish a definitive judgment criterion, we introduce the "Win-Tie-Lose" framework, characterized as follows:

1) Win: InstructionGPT-4 is deemed the winner in two instances, or secures victory once and achieves a draw once; 2) Tie: InstructionGPT-4 achieves a draw twice, or prevails in one instance and succumbs in another; 3) Fail: InstructionGPT-4 faces defeat in two instances, or experiences a loss once and attains a draw once.

The results of this evaluation methodology are depicted in Figure 1. In the context of this figure, Win, Fail, and Tie here denote comparative outcomes when the generation results of InstructionGPT-4 are evaluated against those of MiniGPT-4. Throughout 60 questions, InstructionGPT-4 emerges victorious in 29 instances, experiences failure in 16, and achieves a tie in 15. This evidence underscores the notable superiority of InstructionGPT-4’s response quality in comparison to MiniGPT-4.

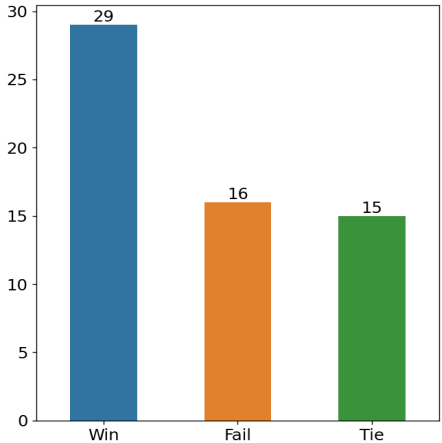


Figure 1: GPT evaluation comparison between InstructionGPT-4 and MiniGPT-4.

### 5.3 Ablation Study

**Analysis of Clustering.** The application of spectral clustering within the data selector mechanism ensures the diversity of the chosen vision-language instruction data. To dissect the contribution of clustering, we conduct an ablation study by excluding the clustering mechanism. The aim here is to assess the role played by clustering in the fine-tuning process. The results of this analysis are presented in the left part of Table 5, highlighting the significance of clustering in enhancing the fine-tuning procedure. By comparing these variations, we elucidate the contributions of clustering and different scoring metrics to the overall efficacy of the data selector methodology. This analysis provides valuable insights into the significance of these components in the data selection process and their subsequent influence on model performance.

**Analysis of Different Scores.** To comprehensively evaluate the impact of distinct scoring metrics on the data selection process, we conduct another ablation study. Each individual scoring metric is isolated and its effect on data selection is scrutinized. As showcased in the right part of Table 5, the models fine-tuned using data selected based on CLIP Score, Answer Length, Reward Score, and GPT Score consistently outperform those generated through random sampling. Notably, the model that emerges from the data selector, an amalgamation of these diverse scoring metrics, demonstrates the highest level of performance attainment.

### 5.4 Demonstrations

In a continued effort to delve into InstructionGPT-4’s proficiency in comprehending visual input and generating reasonable responses, we undertake a comparative assessment of image understanding and conversation abilities between InstructionGPT-4 and MiniGPT-4. This analysis is conducted across a prominent instance of image description and further understanding, outlined in Table 6. InstructionGPT-4 excels in providing comprehensive image descriptions and recognizing intriguing facets within images. In contrast to MiniGPT-4, InstructionGPT-4 showcases a heightened proficiency in identifying text present within images. Notably, InstructionGPT-4 can correctly point out the phrase “Monday, just Monday” on the image. Other instances (Table 7 and Table 8) are shown in the Appendix A.

These instances are meticulously chosen to demand a profound level of image understanding. Despite being fine-tuned using a comparatively small multimodal instruction-following dataset, InstructionGPT-4 remarkably surpasses MiniGPT-4 in terms of its reasoning capabilities within these exemplars. This outcome underscores



Datasets	D.S. w/o clus.	D.S.	Random	CLIP	Reward	Length	GPT
Flickr-30k	19.55 ( $\pm 0.19$ )	<b>20.31</b> ( $\pm 0.07$ )	17.88 ( $\pm 0.26$ )	18.40 ( $\pm 0.03$ )	<b>19.71</b> ( $\pm 0.17$ )	17.90 ( $\pm 0.04$ )	19.14 ( $\pm 0.05$ )
NoCaps	39.74 ( $\pm 0.39$ )	<b>41.63</b> ( $\pm 0.38$ )	38.61 ( $\pm 0.48$ )	40.52 ( $\pm 0.58$ )	<b>40.80</b> ( $\pm 0.09$ )	39.04 ( $\pm 0.26$ )	39.66 ( $\pm 0.14$ )
ScienceQA	23.81 ( $\pm 0.77$ )	<b>24.56</b> ( $\pm 0.39$ )	21.73 ( $\pm 0.45$ )	23.10 ( $\pm 0.69$ )	22.87 ( $\pm 0.54$ )	<b>24.19</b> ( $\pm 0.22$ )	23.77 ( $\pm 0.23$ )
OKVQA	40.44 ( $\pm 0.49$ )	<b>40.69</b> ( $\pm 0.51$ )	36.60 ( $\pm 0.17$ )	36.18 ( $\pm 0.09$ )	<b>39.27</b> ( $\pm 0.29$ )	38.58 ( $\pm 0.28$ )	37.55 ( $\pm 0.36$ )
VSR	<b>40.62</b> ( $\pm 0.34$ )	39.85 ( $\pm 0.31$ )	40.72 ( $\pm 0.72$ )	40.91 ( $\pm 0.28$ )	39.29 ( $\pm 0.25$ )	41.23 ( $\pm 0.10$ )	<b>41.57</b> ( $\pm 0.36$ )
VCR-OC	35.67 ( $\pm 0.92$ )	<b>38.41</b> ( $\pm 0.28$ )	29.67 ( $\pm 1.01$ )	30.49 ( $\pm 0.20$ )	33.22 ( $\pm 0.37$ )	<b>39.84</b> ( $\pm 0.21$ )	30.13 ( $\pm 0.12$ )
VCR-MCI	51.03 ( $\pm 0.91$ )	<b>51.95</b> ( $\pm 0.40$ )	47.17 ( $\pm 0.98$ )	48.22 ( $\pm 0.44$ )	<b>50.11</b> ( $\pm 0.45$ )	49.23 ( $\pm 0.38$ )	48.15 ( $\pm 0.04$ )
Average Score	35.84	<b>36.77</b>	33.20	33.97	35.04	<b>35.72</b>	34.28

Table 5: Ablation study of clustering and different scores. D.S. w/o clus. represents DATA SELECTOR without conducting spectral clustering initially and D.S. simply refers to the standard DATA SELECTOR. The right side includes the performance of models fine-tuned using data filtered by Random selection, CLIP Score, Reward Score, Answer Length, and GPT Score.

InstructionGPT-4’s exceptional aptitude for advanced reasoning, thereby highlighting its heightened capacity for both image comprehension and effective execution of instruction-following tasks.

## 6 Conclusion and Future Work

In this paper, we present a comprehensive analysis of the proposed data selector, evaluating its efficacy in curating valuable instruction data. Our paper culminates with an extensive evaluation of InstructionGPT-4’s performance on multiple benchmark datasets, affirming its superiority in generating contextually coherent and accurate outputs. Ultimately, this research contributes to the broader understanding of instruction tuning and its potential in training efficient multimodal large language models for a myriad of real-world applications.

The central thesis of our research is grounded in the notion that employing a judiciously selected subset of high-quality instruction data can significantly enhance the performance of multimodal large language models. The success of InstructionGPT-4 serves as evidence that inducing instruction data by reasonable selection can lead to significant advancements in the field of multimodal language processing, fostering improved instruction understanding and generation capabilities.

In forthcoming research, extending the analysis to include other models of varying sizes, such as MiniGPT-4-13B and LLaVA, could provide additional insights. Besides, developing a more generalized version of the data selector is another potential avenue for exploration. Additionally, undertaking multimodal instruction mining may provide further insights and unveil new opportunities in this domain. This could enable a broader and more comprehensive understanding of this field.


<div> <div>Demos</div>  </div>	
User	Describe the picture in detail.
InstructionGPT-4	The image is a French bulldog lying on its back on a wooden floor with the caption “Monday, just Monday” written on the top of the image. The dog is sleeping with its eyes closed and its head on its paws. The color scheme is mainly brown with white and black spots on the dog’s fur.
MiniGPT-4	The image is a small dog sleeping on a wooden floor. The dog is brown and has a fluffy fur. The dog is lying on its back with its legs in the air. The caption above the image reads, “Monday just got a little better”.
User	Explain why this meme is funny.
InstructionGPT-4	This meme is funny because it uses a cute and relatable image of a French bulldog to express the common feeling of tiredness or laziness on Mondays. The caption “Monday, just Monday” emphasizes this feeling by making it sound like a statement of fact rather than a complaint. The fact that the dog is sleeping on its back adds to the humor by depicting a situation where the animal is more comfortable than the human. Additionally, the caption is written in all-caps, which gives it a bold and exaggerated effect that makes it more funny. Overall, the combination of the image, caption, and font all contribute to making this meme humorous.
MiniGPT-4	This meme is funny because it is a play on words. The caption reads “Monday just got a little better” which implies that something positive has happened on Monday, but the image shows a sleeping puppy on the floor, which is the opposite of something positive. It is a humorous contrast between the caption and the image. The image of a cute puppy sleeping on the floor is also pleasing to the viewer’s eyes, making the meme more enjoyable.

Table 6: Detailed image descriptions and identifying amusing aspects within images. Compared to MiniGPT-4, InstructionGPT-4 exhibits a superior ability to identify text within images and offers more elaborate descriptions of the visual content. Additionally, InstructionGPT-4 showcases enhanced comprehension of the contextual implications associated with the provided images.

## References

- [1] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [2] OpenAI. Gpt-4 technical report, 2023.
- [3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [4] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [9] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.
- [10] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding, 2023.
- [11] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- [13] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023.
- [14] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.
- [15] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpapasus: Training a better alpaca with fewer data, 2023.

- [16] Openassistant/reward-model-deberta-v3-large-v2. <https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2>, 2023.
- [17] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.
- [18] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022.
- [19] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023.
- [20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [21] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [22] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.
- [23] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*, 2023.
- [24] Yihan Cao, Yanbin Kang, and Lichao Sun. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*, 2023.
- [25] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models, 2023.
- [26] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019.
- [27] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [28] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.
- [29] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023.

# Appendix

## A Additional Demonstrations

Table 7 and Table 8 distinctly indicate that InstructionGPT-4 possesses the capacity to generate responses that are not only more comprehensive but also exhibit a higher level of fluency when compared to those produced by MiniGPT-4.

The story presented by InstructionGPT-4 in Table 7 is vivid and detailed, which involves the key information from the image. Furthermore, when tasked with writing a recipe based solely on a food image in Table 8, InstructionGPT-4 demonstrates a notably superior ability in chain-of-thought by generating robust and effective responses step by step.

These findings collectively emphasize the enhanced language generation prowess of InstructionGPT-4, which contributes to more detailed, expressive, and contextually appropriate responses in various scenarios.


Demos	
	
User	Can you write a story around this image for my child?
InstructionGPT-4	<p>Sure, here's a short story about the animals in the image:</p> <p>One sunny day, a group of animals gathered around a picnic table in a peaceful meadow. The bear, rabbit, and cat were sitting together, sharing a meal of sandwiches, fruit and cookies.</p> <p>The bear was enjoying his sandwich, while the rabbit was nibbling on some grapes. The cat was sipping some milk from a bowl.</p> <p>As they ate, they chatted and laughed, enjoying each other's company. They had a lovely time together, taking in the beautiful scenery around them.</p> <p>After they finished their meal, they decided to take a walk through the meadow. They walked for hours, exploring the different plants and animals they found. They had a great time and felt grateful for the beautiful day and the company of their friends.</p> <p>As they returned to the picnic table, they sat down to rest and share stories of their adventures. They all agreed that it was a perfect day spent in the great outdoors with good friends.</p> <p>The end.</p>
MiniGPT-4	<p>Once upon a time, there was a bear, a rabbit, and a cat who lived in the forest. They spent their days exploring the forest, playing games, and having picnics together. One day, they decided to have a picnic at a table in the middle of the forest. They sat around the table, enjoying their meal and the beautiful scenery around them. The bear ate his share of the food, the rabbit ate her share, and the cat drank some milk. They were all happy and content.</p>

Table 7: Story generation example.

## Demos



User	What are the meals that I can cook with these?
InstructionGPT-4	There are several meals that can be cooked with the ingredients in the refrigerator. For example, you can make a simple smoothie using frozen berries, yogurt, and milk, or a vegetable and fruit salad with a variety of ingredients. You can also make a sandwich or wrap using the ingredients in the fridge. The possibilities are endless, depending on your personal preferences and dietary needs.
MiniGPT-4	This refrigerator contains various types of food, including milk, yogurt, cereal, fruits, and vegetables.
User	I would like to do the fruit salad. Show me the recipe.
InstructionGPT-4	<p>Here is a simple recipe for a fruit salad using ingredients from the refrigerator:</p> <p>Ingredients:</p> <ul style="list-style-type: none"> <li>* 1 cup mixed fruit (e.g., berries, cherries, peaches, plums, etc.)</li> <li>* 1/4 cup plain Greek yogurt</li> <li>* 1 tablespoon honey or maple syrup</li> <li>* 1 tablespoon lemon juice</li> <li>* 1/4 cup chopped nuts (e.g., almonds, walnuts, cashews, etc.)</li> <li>* 1/4 cup dried cranberries or raisins</li> </ul> <p>Instructions:</p> <ol style="list-style-type: none"> <li>1. In a large bowl, combine the mixed fruit, Greek yogurt, honey or maple syrup, and lemon juice.</li> <li>2. Stir to combine the ingredients.</li> <li>3. Add the chopped nuts and dried cranberries to the bowl.</li> <li>4. Stir to combine the ingredients.</li> <li>5. Serve the fruit salad in bowls, garnished with chopped nuts and dried cranberries if desired.</li> </ol> <p>Enjoy your delicious fruit salad!</p>
MiniGPT-4	<p>Sure, here's a recipe for a simple fruit salad:</p> <p>Ingredients:</p> <ul style="list-style-type: none"> <li>* Bananas, peeled and sliced</li> <li>* Apples, peeled and sliced</li> <li>* Grapes</li> <li>* Strawberries</li> <li>* Blueberries</li> <li>* Rice</li> <li>* Sugar</li> <li>* Lemon juice</li> </ul> <p>Instructions:</p> <ol style="list-style-type: none"> <li>1. In a large bowl, mix together the sliced bananas, apples, grapes, strawberries, blueberries, and rice.</li> <li>2. In a separate bowl, mix together the sugar and lemon juice until the sugar is fully dissolved.</li> <li>3. Pour the sugar mixture over the fruit mixture and toss to coat everything evenly.</li> <li>4. Serve the fruit salad cold or chilled, garnished with lemon wedges or mint leaves if desired.</li> </ol> <p>That's it! This simple fruit salad is a great way to get in some healthy fruits and veggies while satisfying your sweet tooth. Enjoy!</p>

Table 8: Generating chain-of-thought responses.



## B GPT Prompt

We provide the detailed prompt to GPT-4 used for rating scores in Table 9. It is similar to the prompt for rating and filtering training data in Alpargatus [15].

GPT Prompt	
System Prompt	We would like to request your feedback on the performance of an AI assistant. The assistant provides a caption based on an image and an instruction. Instruction: [Instruction] Caption: [Caption]
User Prompt	Please rate according to the quality and variety of the caption to the instruction. Each assistant receives a score on a scale of 0 to 100, where a higher score indicates higher level of the quality and variety. Please first output a single line containing the value indicating the scores. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias. The instruction and caption are displayed following without image.

Table 9: Prompt  $p_G$  to GPT-4 for rating multimodal data.