Rethinking Overlooked Aspects in Vision-Language Models

Yuan Liu, Le Tian, Xiao Zhou, Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, China
{bensenliu, letian, chappyzhou, withtomzhou}@tencent.com

Abstract

In recent years, large vision-language models (LVLMs), such as GPT4-V, have advanced significantly, primarily due to the transformative impact of large language models. Among these LVLMs, LLaVA stands out as a widely adopted model, offering several key advantages: (i) Simplicity—LLaVA consists of three main components: a vision encoder, a lightweight adapter (e.g., MLP), and a large language model (LLM). This modular architecture facilitates the integration of state-of-the-art (SOTA) models into any component to boost performance. (ii) Efficiency—With a modest pre-training dataset of 558k images and 665k instances of supervised fine-tuning data, LLaVA-1.5-13B achieves impressive results across numerous benchmarks. Subsequent works have begun to (i) incorporate substantially more data during pre-training, and (ii) utilize a more diverse and larger instruction tuning dataset. In this report, we aim to investigate two important, yet previously overlooked, aspects: (i) ***the efficiency of data during pre-training****—whether the model's performance consistently improves with the addition of more pre-training data; (ii)* ***how to choose instruction tuning datasets****—the effectiveness of the SFT datasets used in existing works and the methodology for selecting the most impactful ones. Through a meticulous and comprehensive examination, we have discovered that a naive increase in the size of the pre-training dataset does not effectively enhance the performance of Vision-Language Models (VLM). In fact, it may even lead to a degradation in performance. Regarding the SFT data, we have developed an effective pipeline to identify the most efficient SFT dataset. Our study reveals that not all SFT data employed in existing works are necessary and can be optimized for enhanced performance. With the findings of this report, we hope to encourage future research to focus more on the data used during pre-training and supervised fine-tuning to further push the boundaries of vision-language models.*

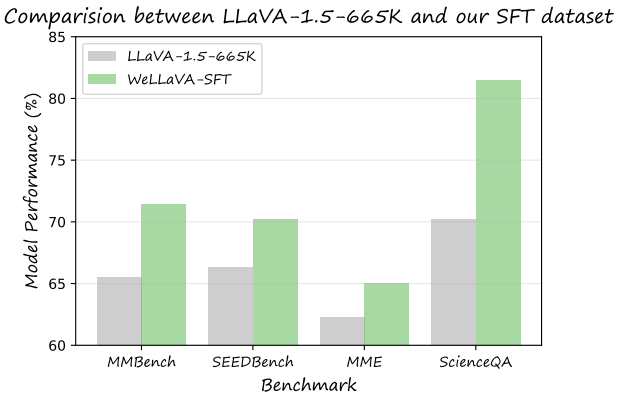

Figure 1. **Performance of LLaVA-1.5 using LLaVA-1.5-665K and our SFT dataset.** We have developed a strategy, termed **Individual Select**, which is designed to select the most effective datasets from a plethora of publicly available SFT datasets. The LLaVA-1.5 model, fine-tuned with Vicuna-7B[13] on the final composition of SFT datasets that we have obtained, yields substantial improvements compared to the baseline. The original MME scores are mapped to a range of 0 to 100.

1. Introduction

Large language models (LLMs) have achieved significant progress in recent years. Some models, such as GPT-4 [1] and Claude3, have reached or even surpassed human performance in various aspects. Compared to LLMs, large vision-language models (LVLMs) can solve much more complex problems that the original text-only LLMs could not, such as image understanding and question-answering. With the rapid development of LLMs, LVLMs have also demonstrated remarkable achievements. Models like the recent GPT-4o, Qwen-VL-Max, and Step-1V show promising capabilities in solving increasingly complex image-related problems, including geometry matching and optical character recognition. However, all these models are proprietary, and the details behind them are not publicly known. Despite the existing gap between open-source models and these proprietary models, recent progress [12, 36] has been made, which is gradually narrowing this divide.

We find that the latest advancements in visual language

models are largely driven by data, including pre-training data and instruction tuning (SFT) data. For instance, models like InternVL-1.5[12], Qwen-VL-Max[3], and DeepSeek-VL[41] utilize web-scale pre-training datasets such as Laion-5B[54] and COYO[5], enabling them to reach a pre-training data volume of 1B. Simultaneously, compared to previous works, their instruction tuning datasets are not only larger in scale but also richer in diversity. For example, InternVL-1.5 divides the sft dataset into 11 subclasses and collects corresponding open-source datasets for each subclass, a practice also adopted by DeepSeek-VL. For **pre-training datasets**, there exists a scaling law in the LLM field[22], which suggests that as the model size increases and the pre-training dataset size is concurrently expanded, the model's performance will also increase synchronously. However, no work in the vision-language model field has yet conducted a comprehensive experiment of this nature. Therefore, we are curious: 1) For the same model, if we continuously increase the amount of pre-training data using existing open-source multimodal datasets, will the model's performance also grow synchronously? 2) When we enlarge the model size and concurrently increase the pre-training data volume, will visual language models also demonstrate a scaling law similar to that in LLM? Regarding the **SFT datasets**, for a long time, everyone has been conducting instruction-following training based on the dataset proposed in LLaVA-1.5[35]. Compared to the data in LLaVA-1.5, the latest works have introduced datasets with more categories and larger quantities. However, given the varying quality of these newly introduced datasets and the significant overlap between different datasets, it raises the question of whether these datasets all play a key role in enhancing the model's general capabilities.

LLaVA-1.5 is a very simple and efficient model, mainly composed of three parts: a vision encoder, a vision-language adapter, and a large language model. Upon its introduction, it receives significant attention and has been used and improved upon in many subsequent works. We designed several experiments based on LLaVA-1.5, the simplest model, hoping to reveal answers to the above questions. Firstly, regarding the pre-training dataset, we extracted seven sets of data from LAION-5B-en, with sizes ranging from 1M to 100M, and trained the same model on these datasets. Simultaneously, to observe whether the model's performance steadily improves with the increase in model size and data volume, we select Vicuna-7B/13B, Qwen1.5-Chat-7B/14B, and Yi-Chat-6B/34B to study this phenomenon. As for the SFT dataset, we used the dataset in LLaVA-1.5 as the base version. Referring to the taxonomy of the SFT dataset in InternVL-1.5[12], we proposed a method called **Individual Select**, which selects the most effective dataset in each category on a single dataset granularity. Through a large number of experiments, we have found that:

(1) When we naively use existing datasets to increase the amount of pre-training data for the model, the performance of the model does not improve, and may even cause a decline in the model's performance.

(2) For the same type of models, such as Vicuna-7B and Vicuna-13B, increasing the model size while also increasing the amount of pre-training data in parallel, the ultimate improvement in model performance is largely due to the use of a larger model.

(3) The SFT dataset used in the latest work has a lot of redundancy, and there is still a large exploration space.

In summary, the aim of this report is not to propose a state-of-the-art (SoTA) model, but rather to explore crucial yet previously overlooked facets of vision-language development and research. We anticipate that the findings from this report will offer valuable insights and guidance for the future advancement of vision-language models.

2. Related Works

Pre-training Pre-training is a prevalent technique across various domains, such as computer vision and natural language processing. The primary objective of pre-training is to equip a randomly initialized model with the capability to accomplish general tasks. For instance, in computer vision, tasks like classification and instance segmentation often benefit from pre-training the model using methods like masked image modeling[21, 39, 40]. Moreover, the recent advancements in large language models can be attributed to extensive pre-training on vast datasets. In the context of vision-language models (VLMs), they typically comprise three components: a vision encoder, an adapter, and a language model. Given that the vision encoder and language model are trained separately, they exhibit different feature distributions. Therefore, the primary focus of a vision-language model is on vision-language alignment. During this alignment process, some studies[9, 10, 35] opt to freeze both the vision encoder and language model, training only the lightweight adapter with a smaller dataset. Conversely, others choose to freeze only the language model, training both the vision encoder and adapter with a larger dataset. However, there exists no agreement about how much data should be used during pre-training.

Instruction Tuning Instruction tuning is a crucial technique in both large language models (LLMs) and vision-language models (VLMs), enabling the model to accurately follow human instructions. However, research on instruction tuning in VLMs significantly lags behind that in LLMs. LLaVA[17], a pioneering work in this field, suggests using text-only GPT4 to generate a large number of instruction tuning datasets. The instruction tuning dataset in LLaVA

primarily comprises three categories: conversation, complex reasoning, and detailed description. Subsequently, InstructBLIP[16] proposed the use of a vast amount of academic datasets as instruction tuning datasets and designed specific prompts for each dataset. Models fine-tuned on these datasets demonstrated impressive performance at the time. Following this, a series of recent works[11, 12, 18, 35] began to incorporate datasets from various sources, such as those generated by GPT-4V and academic datasets, further elevating the performance of vision-language models to unprecedented levels. However, it is unfortunate that, to date, no study has provided insights into the type of SFT dataset that is most efficient, or whether we need all the datasets to fine-tune our model.

3. Overlooked Aspects

3.1. Creating a Stable Baseline

Benchmarks Selection. We anticipate that the model will excel in a variety of general tasks, rather than concentrating on a single specific task. Consequently, we expect a benchmark should be able to evaluate different aspects of a model comprehensively. Recently proposed benchmarks such as MMBench[38], MME[19], and SEED-Bench are designed to provide a thorough evaluation of a model's performance. In comparison to earlier benchmarks[20, 23], these cover a wider range of ability dimensions, offering a more comprehensive insight into the capabilities of the evaluated model. Therefore, we choose these benchmarks as our guidelines for setting selection during our exploration. Furthermore, considering the frequent appearance of scientific problems in our daily lives, we have also included the metric from ScienceQA[44] in our guidelines. All the subsequent results are obtained using the evaluation toolkit, VLMEvalKit[1].

Model Training Framework In order to optimize training efficiency, we employ an in-house training framework developed by WeChat. This framework improves the data loading pipeline by introducing a more efficient data format. To further enhance training efficiency, we concatenate multiple samples to achieve a maximum sequence length of 4096, which is subsequently fed into the model. This method is consistent with the standard practices used in training large language models. Furthermore, our framework supports various types of model parallelism, such as tensor parallel[28] and pipeline parallel, as well as data parallelism. Before embarking on a comprehensive exploration, it is essential to train a baseline model, for instance, LLaVA-1.5[35], to validate the accuracy of our training framework. We maintain consistency with LLaVA-1.5 in terms of all datasets, including pre-training and supervised fine-tuning, as well as hyper-parameters. Additionally, we

[1] https://github.com/open-compass/VLMEvalKit

LLM	MME[19]	MMB-dev[38]	SQA^I[44]	$SEED^I$[29]
Vicuna-7B[13]	1808.4	65.2	66.8	65.8
Vicuna-7B[13]	1772.2 (-36.2)	64.1 (-1.1)	70.0 (+3.2)	65.1 (-0.7)
Yi-6B[63]	1772.8	70.4	73.5	68.6
Yi-34B[63]	1840.3	74.8	75.2	72.0
Qwen-1.5-7B[3]	1657.4	67.4	67.0	66.6
Qwen-1.5-14B[3]	1801.7	70.6	71.6	68.2

Table 1. **Comparison with the official implementation of LLaVA-1.5.** MMB-dev: the *dev* set of MMBench. SQA^I[44]: the image split of ScienceQA. $SEED^I$[29]: the image split of Seed-Bench. The gray line illustrates the model's performance as per the official implementation, while the green line represents the model's performance achieved through our training framework.

lr^v	lr^a	MME	MMB-dev	SQA^I	$SEED^I$
N/A	1e-3	1772.2	64.1	70.0	65.1
2e-5	2e-4	1744.0.0 (-32.2)	65.5 (+1.1)	70.2 (+0.2)	66.3(+1.2)

Table 2. **Improved pre-training settings.** Unfreezing the vision encoder is beneficial to improve the performance of LLaVA. lr^v: learning rate for vision encoder. lr^a: learning rate for the MLP adapter. N/A: fix the vision encoder.

select several language models, such as Vicuna[13], Qwen-1.5[3], and Nous-Hermes-2-Yi[63], to ensure our conclusions are more generalizable and convincing. Just as shown in Table 1, the model of our implementation is comparable to that of the official implementation. When the language model is replaced, models obtained by our training frameworks can also abtain reasonable results. The official implementation results are primarily sourced from the OpenCompass leaderboard[2][15], with the exception of the ScienceQA results, which are referenced from the original paper.

Improved Pre-training Settings The original LLaVA fixes the vision encoder, focusing solely on the MLP adapter's pre-training to enhance efficiency. However, an increasing number of studies[10, 12, 18] suggest that jointly training the vision encoder and the adapter can be advantageous. This approach allows for the adjustment of the feature distribution to the generation task, thereby enhancing the vision encoder's feature extraction capabilities. In our work, we also unfreeze the vision encoder and assign different learning rates to the vision encoder and the MLP adapter. As demonstrated in Table 2, this configuration improves LLaVA's performance non-trivially.

3.2. Scaling Up the Pre-training Data

Rather than solely relying on the 585K data from LLaVA for pre-training, a growing body of research is incorporating significantly larger datasets during this phase. For instance, Qwen-VL[3] and InternVL[11, 12] utilize web-scale

[2] https://rank.opencompass.org.cn/leaderboard-multimodal

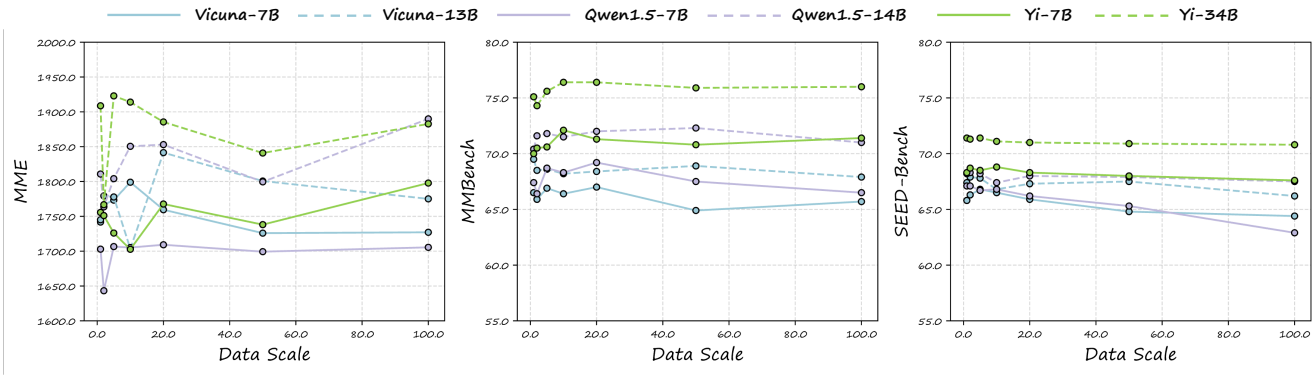

Figure 2. **Pre-training data scaling law.** We investigated this phenomenon in large vision-language models using three different types of mainstream Large Language Models (LLMs). As we increased the size of the pre-training dataset from 1 million to 100 million samples, the model's performance remained nearly consistent, with some instances of degradation observed.

pre-training datasets such as LAION-5B[54] and COYO-700M[5]. While these studies have shown promising results, a more thorough analysis is required to fully comprehend data efficiency during the vision-language pre-training stage. In this section, we delve further into this topic by training a model on a subset of LAION-5B, ranging from 1M to 100M. Additionally, we investigate the relationship between the model size and the scale of the pre-training dataset. To strengthen our conclusions, we also include three types of large language models, namely Vicuna[13], Qwen-1.5[3], and Hous-Hermes-2-Yi[63], in our experiments. The detailed results are depicted in Figure 2.

Three main insights can be gleaned from the trends in Figure 2: 1) The pre-training datasets currently used in vision-language pre-training are quite inefficient. As we scale up the pre-training dataset from 1M to 100M, we only observe marginal improvements across the three benchmarks, and performance even deteriorates as we increase the dataset size beyond 50M. For example, the performance of Qwen1.5-7B on SEED-Bench drops 3.3 points as we scale the size of pre-training data from 20M to 100M. 2) Scaling up the size of the LLM can yield substantial improvements. For instance, Vicuna-13B outperforms Vicuna-7B by 3.0 when 1M data is used for pre-training. 3) The performance trend for different model sizes is almost consistent across different pre-training data scales. This observation contrasts with the trend in large language models, which suggests that larger models using more data can obtain further improvement.

Based on the observed experimental phenomena, it is evident that simply scaling up the size of the pre-training dataset does not effectively enhance the performance of vision-language models. A more promising approach lies in enhancing the quality and diversity of the data, as suggested by studies on LLM [6, 59], which emphasizes the importance of utilizing more than just generic image-caption pairs.

3.3. Instruction Dataset Selection

Based on the taxonomy of SFT data in InternVL-1.5[12], we have added a new category related to screenshots, resulting in a total of 12 categories. For the datasets in each category, we still use the datasets given in InternVL-1.5 as the base version, and then introduce datasets used in other works, such as DeepSeek-VL[41]. Before conducting detailed ablation experiments on the datasets, we conduct a comprehensive study of the selected datasets and eliminated some of the lower quality ones. At the same time, we find that some datasets in InternVL-1.5 are not accurately classified. For example, ALLaVA[9] contains both picture caption data and conversation data, but InternVL-1.5 completely places it in the conversation category. Through filtering and reclassification of the datasets, we obtain a base version of the datasets before further selection (Table 3).

Given that LLaVA-1.5-665K is currently the most extensively utilized dataset for visual instruction tuning, we select it as the starting point for our investigation. Moreover, as demonstrated by previous studies, substituting the *Detailed Description* data in LLaVA-1.5-665K with data from ShareGPT4V can yield further enhancements. Consequently, we opt for the improved version of LLaVA-1.5-665K as our ultimate starting point. The **Individual Select** strategy then selects the most effective datasets from Figure 3 and incorporates them into LLaVA-1.5-665K. The workflow of the **Individual Select** strategy are as follows:

(1) For each dataset (candidate) from a category in Table 3, we incorporate it into the baseline dataset and fine-tune the model on this newly constructed dataset.

(2) If the model's performance surpasses or is comparable to that achieved when trained on the baseline dataset, we include the candidate dataset in the candidate pool. If not,

Task	Dataset
Captioning	ShareGPT4V [10], LAION-GPT4V, TextOCR-GPT4V [8] SVIT(cap) [66], ALLaVA(cap) [9], LVIS-Instruct4V(cap) [60]
General QA	VSR [32], IConQA [43]
Science	AI2D [25], ScienceQA [44], TQA [26]
Chart	ChartQA [48], MMC-Inst[34], DVQA (en)[24], PlotQA [51], UReader [62]
Mathematics	GeoQA+ [7], TabMWP (en) [46] CLEVR-Math/Super [30, 31], Geometry3K[42]
Knowledge	KVQA [53]
OCR	InfoVQA [50], TextVQA [55], ArT [14] SynthDoG [27], ST- VQA [4]
Document	DocVQA [49],
Grounding	RefCOCO+/g [47, 64]
Conversation	ALLaVA(conv) [9], LVIS-Instruct4V(conv) [60] SVIT(conv) [66], LLaVAR [65], VisualDialog [17]
Text-only	OpenHermes2.5 [58], Alpaca-GPT4 [57], LIMA [67]
Screen	ScreenQA [2]

Table 3. **Base version of SFT datasets to be ablated.** We utilize the **Individual Select** strategy to incrementally select the most effective datasets from the aforementioned ones, and incorporate them into LLaVA-1.5-665K to enhance the richness of the SFT dataset. cap: the caption split. conv: the conversation split.

we discard it. Ultimately, we concatenate all the datasets from the candidate pool and integrate them into the baseline dataset to establish a new baseline dataset.

(3) We then employ this new baseline dataset, iterate through all the subsequent categories, and repeat Steps 1 and 2 above.

In Step 2, as outlined above, we aim to avoid overfitting the four benchmark datasets. Therefore, if the model trained on the newly constructed dataset yields performance that is merely comparable to, and does not exceed, that of the model trained on the baseline dataset, we still include it in the candidate pool. Furthermore, we have observed that if two datasets individually contribute to improvements, their combination can lead to even further enhancements. Figure 3 shows the details of **Individual Select** and the final datasets we choose to be added to the improved version of LLaVA-1.5-665K. he final selection of datasets we use comprises nine categories, reducing the original number from 37 to 17.

4. Experiments

Experiment Setting Apart from the modifications introduced in Section 3, all other settings remain consistent with those of LLaVA-1.5. Specifically, we employ OpenAI's CLIP-Large-336px[52] as the vision encoder. The learning rate is linearly warmed up during the initial 3% of iterations, after which a cosine decay learning rate strategy is implemented.

4.1. Comparison with Other Models

In this section, we provide a thorough comparison of our model, which utilizes techniques from Section 3, with other state-of-the-art (SOTA) models across seven benchmarks, namely MMBench[38], MME[19], MathVista[45], HallusionBench[33], SEEDBench[29], and LLaVABench[37]. We have chosen two models for this comparison: Vicuna-1.5-7B/13B. For both models, we select the best pre-trained versions as identified by the ablation study illustrated in Figure 2. The detailed results are presented in Table 4. As the table indicates, when comparing our model with other models using the same LLM, such as Vicuna-7B/13B, our model outperforms the others by a significant margin overall. Despite being directly based on LLaVA-1.5, our model even surpasses LLaVA-Next, which introduces new strategies like the use of high-resolution images. This substantial performance improvement further underscores the importance of exploring the composition of SFT datasets.

4.2. Ablation Studies

Meticulously choose the SFT datasets is important. In this section, we compare the models fine-tuned on all these datasets from Table 3 and datasets obtained by **Individual Select**. Just as shown, indiscriminately fine-tuning the model on the all datasets will not bring improvements, or even degrade the performance. Another drawbacks of fine-tuning the model on all the datasets is the heavy computation overhead, since the size of these datasets is about six times of ours.

Consistent improvements. The paragraph, as depicted in Figure 4, demonstrates a consistent overall improvement during the dataset selection process as more datasets from each category are incorporated. We also observe a nearly linear improvement trend for each benchmark, with the exception of MME. However, the performance trend of MME remains relatively stable after the inclusion of datasets from the *Caption* category. Notably, there is a significant improvement in ScienceQA following the introduction of datasets from the *Science* category. This can be attributed to the addition of the *train split* of ScienceQA. This observation underscores the potential for enhancing model performance on a specific task by introducing a dataset with a similar distribution to that task. However, it is crucial to maintain a balance with other datasets to prevent a decline in the model's general capabilities.

5. Discussion, Limitation and Future Works

Discussion As discussed in Section 3, we find that simply increasing the size of the pre-training dataset does not consistently yield improvements. We propose two possible

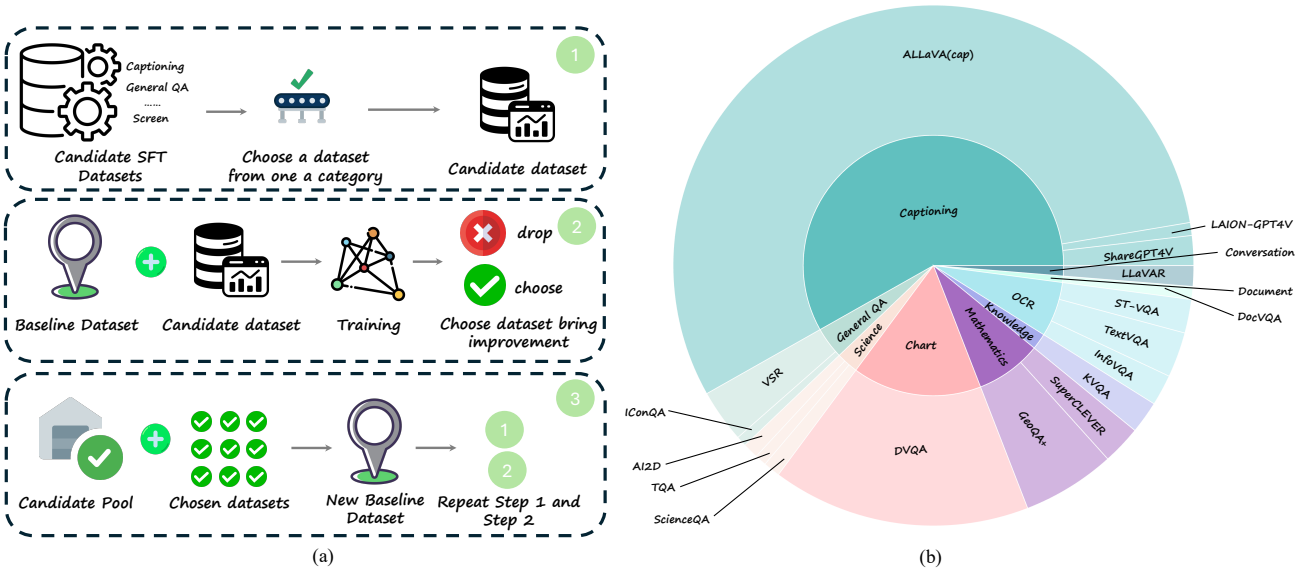

Figure 3. **The workflow of Individual Select (a) and the SFT datasets we finally choose to be added to the baseline dataset.**

Method	LLM	Vision	MMBench (en)	MMBench (cn)	MME	MathVista	HallusionBench	SEED^I	LLaVABench
LLaVA-1.5[35]	Vicuna-7B[13]	CL[52]	64.3	58.3	1808.4	25.6	27.6	66.1	65.4
LLaVA-Next[36]	Vicuna-7B[13]	CL[52]	69.2	62.3	1769.1	31.5	27.6	69.6	72.7
CogVLM[61]	Vicuna-7B[13]	E2CL-E[56]	65.8	55.9	1736.6	35.0	**35.4**	68.8	**73.9**
Ours	Vicuna-7B[13]	CL[52]	**72.6**	**65.8**	**1818.7**	**42.6**	31.9	**69.9**	62.7
LLaVA-1.5[35]	Vicuna-13B[13]	CL[52]	67.7	63.6	1780.8	27.7	24.5	68.2	66.1
LLaVA-Next[36]	Vicuna-13B[13]	CL[52]	68.8	61.9	1745.6	34.1	**31.8**	70.1	**73.9**
ShareGPT4V[10]	Vicuna-13B[13]	CL[52]	69.8	65.1	1853.1	29.3	28.4	**70.6**	69.1
Ours	Vicuna-13B[13]	CL[52]	**74.9**	**69.8**	**1879.8**	**39.1**	31.4	70.3	65.2

Table 4. **Comparison with other methods across benchmarks for instruction-following LMMs.** We select the evaluation metric for each method based on its original paper, if available. If not, we use the metric provided by the leaderboard of VLMEvalKit. Vision: vision encoder. CL: OpenAI CLIP-L-336px. E2CL-E: EVA2-CLIP-E.

Dataset	LLM	MME	MMB-dev	SQA^I	SEED^I
All	Vicuna-7B	1790.0	70.5	80.1	70.0
Ours	Vicuna-7B	1818.7 (+28.7)	73.0 (+2.5)	81.6 (+1.5)	69.9 (-0.1)

Table 5. **Model performance comparison between fine-tuned on all datasets and datasets we select.** All: all datasets from Table 3.

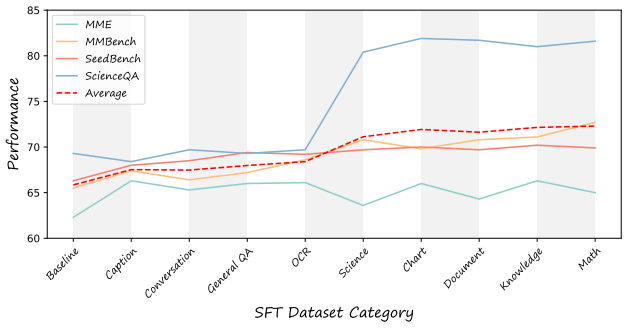

Figure 4. **We observe consistent improvements as we incorporate these selected datasets from each category.** The original MME scores are mapped to a range of 0 to 100. Average: the average score of the four metrics.

reasons for this: i) The quality of the pre-training dataset may be suboptimal. LAION-5B, which is crawled from the Internet and only subjected to basic data filtering processes such as image-text similarity filtering, may contain a significant amount of noise, including grammatical errors in text and incorrect punctuation. ii) The vision encoder has already been pre-trained on a dataset with a distribution similar to that used in the vision-language alignment pre-training. Given that the vision encoder is not frozen, the vision-language pre-training focuses on vision-language alignment and the injection of new knowledge into the vision encoder. This new knowledge injection involves instilling abilities that were not learned during the vision encoder pre-training. Therefore, using more data with a distribution similar to that used for vision encoder pre-training does

not result in substantial improvements during the vision-language pre-training stage.

Limitation In our exploration of SFT datasets, we try to prevent overfitting. The model, fine-tuned on our dataset, also performs well on other benchmarks as shown in Table 4. However, the performance gap on other benchmarks is smaller than that on the three benchmarks used in the ablation study, suggesting that more appropriate benchmarks for ablation should be selected. Currently, our investigation of the SFT dataset is primarily conducted at the level of individual datasets. While this approach has led to significant improvements in model performance, it is relatively rudimentary. Furthermore, we have not yet explored aspects related to the quality and distribution of the SFT dataset in our current work.

Future Works The primary objective of this paper is not to introduce a state-of-the-art (SoTA) model that performs competitively across a variety of benchmarks. Instead, our focus is on investigating some aspects that have been previously overlooked. While the model we have developed performs satisfactorily on a series of benchmarks, there is room for further refinement in terms of user experience. In future work, we will concentrate on three main areas: i) Dataset quality: This includes both pre-training and SFT datasets. We plan to undertake a series of explorations on how to clean existing datasets and generate more effective ones. ii) Knowledge injection: We aim to move beyond solely relying on general image-text caption datasets. Our intention is to incorporate datasets with different distributions during pre-training, such as Optical Character Recognition (OCR). iii) Incorporation of recent techniques in Vision-Language Models (VLM): This includes the use of high-resolution images.

6. Conclusion

In this study, we delve into critical yet previously neglected aspects of vision-language models, such as the scaling law during pre-training and the selection of the most effective dataset for instruction fine-tuning. We consider three types of Language Learning Models (LLMs) and assemble seven splits of pre-training datasets, with the total count ranging from 1M to 100M. Our extensive experiments reveal that simply increasing the size of the pre-training dataset does not necessarily yield significant improvements and may even degrade the model's performance. Furthermore, we introduce a strategy, termed **Individual Select**, to identify the most effective datasets from a vast pool of publicly available candidates. This approach leads us to an effective composition of instruction tuning (SFT) datasets. Models fine-tuned on these datasets demonstrate substantial improvements over the baseline and outperform models that are indiscriminately fine-tuned on all SFT datasets. This underscores the need for a more thoughtful composition of SFT datasets.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. Screenai: A vision-language model for ui and infographics understanding, 2024. 5

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3, 4

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 5

[5] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2, 4

[6] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 4

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

[8] Jimmy Carter. Textocr-gpt4v. https://huggingface.co/datasets/jimmycarter/textocr-gpt4v, 2024. 5

[9] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 2, 4, 5

[10] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2, 3, 5, 6

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and

aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 3

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 2, 3, 4

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1, 3, 4, 6

[14] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 5

[15] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 3

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[17] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2, 5

[18] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 3

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 3, 5

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[24] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018. 5

[25] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 5

[26] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017. 5

[27] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[28] Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5, 2023. 3

[29] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3, 5

[30] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023. 5

[31] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 5

[32] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 5

[33] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 5

[34] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong

Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 5

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 3, 6

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 6

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[38] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 3, 5

[39] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*, 2023. 2

[40] Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5361–5372, 2023. 2

[41] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 4

[42] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 5

[43] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 5

[44] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3, 5

[45] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 5

[46] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023. 5

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5

[48] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[49] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 5

[50] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 5

[51] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 5

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 6

[53] Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, 2019. 5

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4

[55] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 5

[56] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 6

[57] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 5

[58] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. 5

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov,

Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4

[60] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 5

[61] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 6

[62] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model, 2023. 5

[63] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 3, 4

[64] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5

[65] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 5

[66] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 5

[67] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 5