ShareGPT4V: Improving Large Multi-Modal Models with Better Captions

Lin Chen*[1,3], Jinsong Li*[2,3], Xiaoyi Dong[2,3], Pan Zhang[3], Conghui He[3], Jiaqi Wang[3], Feng Zhao†[1], Dahua Lin†[2,3,4]

[1]MoE Key Laboratory of BIPC, University of Science and Technology of China
[2]The Chinese University of Hong Kong [3]Shanghai AI Laboratory
[4]Centre for Perceptual and Interactive Intelligence (CPII)

Abstract. Modality alignment serves as the cornerstone for large multi-modal models (LMMs). However, the impact of different attributes (e.g., data type, quality, and scale) of training data on facilitating effective alignment is still under-explored. In this paper, we delve into the influence of training data on LMMs, uncovering three pivotal findings: 1) Highly detailed captions enable more nuanced vision-language alignment, significantly boosting the performance of LMMs in diverse benchmarks, surpassing outcomes from brief captions or VQA data; 2) Cutting-edge LMMs can be close to the captioning capability of costly human annotators, and open-source LMMs could reach similar quality after lightweight fine-tuning; 3) The performance of LMMs scales with the number of detailed captions, exhibiting remarkable improvements across a range from thousands to millions of captions. Drawing from these findings, we introduce the ShareGPT4V series for advanced modality alignment. It includes ShareGPT4V, consisting of 100K high-quality captions curated from GPT4-Vision; ShareGPT4V-PT, containing 1.2M captions produced by our Share-Captioner that can be close to the captioning capabilities of GPT4-Vision; and ShareGPT4V-7B, a simple yet superior LMM excelling in most multi-modal benchmarks, which realized better alignment based on our large-scale high-quality captions. The project is available at `https://sharegpt4v.github.io/`.

Keywords: Large Multi-modal Models · Modality Alignment · High-quality Captions

1 Introduction

Recent breakthroughs in artificial intelligence have been driven notably by the development of large language models (LLMs) [2,4,10,11,14,57,62]. Following the evolution, modality unification via LLMs becomes the inevitable tendency, and visual-aligned multi-modal LLMs [3, 5, 12, 32, 33, 37, 63, 63, 67–69] have witnessed ever-changing advances in recent days. Putting aside the difference in detailed implementation, multi-modal data from diverse sources are collected to facilitate

* Equal contribution † Corresponding authors.

Table 1: Comparison of data type and quality. LLaVA-1.5-7B serves as the baseline. We collect 23K captions from GPT4 [33], BLIP2 [28], GPT4-Vision [42] and 23K VQA data [60] from GPT4-Vision and integrate these data into baseline's SFT data to investigate the impact of data type and quality. **Note that the first row is just the original baseline without data modification.**

Data sources	MME	MMB	QBench	VizWiz	SEED	MM-Vet	LLaVA
GPT4 Caption-23K [33]	1510.7	64.3	58.7	50.0	66.2	30.5	63.4
BLIP2 Caption-23K [28]	1453.9	63.7	57.9	50.2	65.7	28.8	61.5
GPT4-Vision VQA-23K [60]	1494.4	65.0	61.0	50.9	**66.9**	31.8	67.7
GPT4-Vision Caption-23K [42]	**1516.9**	**65.3**	**62.0**	**51.3**	66.8	**34.0**	**71.6**

the alignment between the new modality and language, resulting in the large multi-modal models (LMMs) that have amazing capability in the new modality.

Despite their achievements, the vast data usage difference among them hinders the community from understanding the actual influence of different training data on the modality alignment, including the data type, quality, and scale.

In this paper, we focus on this foundation problem and start with a series of analyses. As depicted in Table 1, we substitute the captions in LLaVA-1.5 with data from various sources and report the results. When comparing with captions imagined by GPT4 [41] or brief captions generated by BLIP2 [28], the comprehensive and precise captions annotated by GPT4-Vision aid the LMM in achieving superior modality alignment, demonstrating excellent results across all benchmarks. Interestingly, even when compared with high-quality VQA data generated by GPT4-Vision, the LMM trained with detailed captions still performs better on all VQA benchmarks. This finding indicates that instead of further accumulating homogeneous VQA data, integrating detailed captions into the SFT data can bring more performance gains to the LMMs. We posit that these detailed captions, which provide a dense and precise correspondence between vision and language, are crucial for efficient modality alignment.

Given the significant improvements achieved with high-quality captions, a natural question arises: Could superior captions from professional human annotators further enhance the performance of LMMs? Interestingly, we find that the performance enhancement between captions from human annotators and state-of-the-art LMMs, such as GPT4-Vision [42] and Gemini Pro [55], is comparable (seen in Figure 5(a)). We contend that while these LMMs may still exhibit occasional hallucinations, they frequently generate more comprehensive descriptions of image content than humans, thereby establishing more visual-language correspondences for improved modality alignment. Moreover, we discover that open-source LMMs could attain a similar captioning capability with lightweight fine-tuning, enabling the cost-effective scaling of high-quality captions.

Besides quality, the quantity of training data is also a crucial factor linked to the scalability of LMMs intimately. Scalability, the ability to continually improve with an increase in training data, is an important and already proven capability of the pure large language models. However, the scalability of LMMs on multi-modal data remains unclear. To investigate this, we expand our high-quality captions from the thousands to the millions using a meticulously constructed pipeline. This process leads to a notable improvement in performance as the volume of data increases, suggesting that the modality alignment of LMMs using high-quality caption data is indeed scalable.

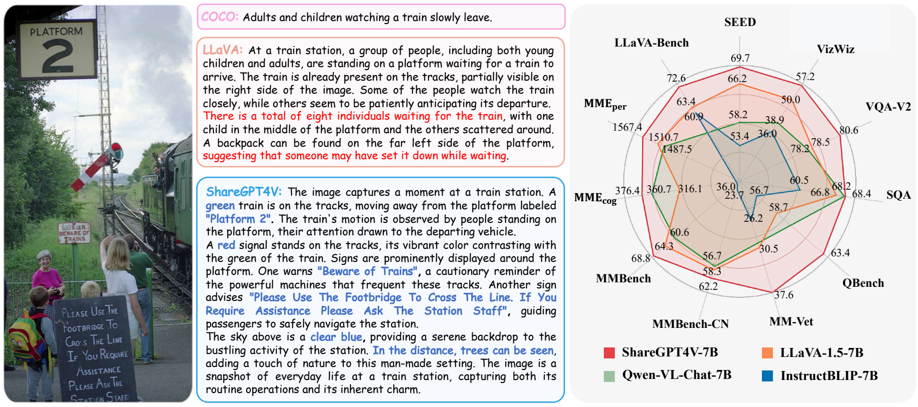
In practice, we first curate images from diverse sources and employ data-specific prompts for GPT4-Vision to generate 100K detailed captions, namely the ShareGPT4V dataset. We showcase a comparison between the caption in our dataset and those utilized by recent LMMs in Figure 1(a). These captions, averaging **942 characters**, encompass a comprehensive range of image information, such as world knowledge, object properties, spatial relation, aesthetic evaluation, etc. Following this, we conduct lightweight fine-tuning on an open-source LMM with the ShareGPT4V dataset, obtaining Share-Captioner, a model capable of generating captions as comprehensive as GPT4-Vision. We then methodically utilize our Share-Captioner to expand our detailed caption dataset to 1.2 million for pre-training, forming the ShareGPT4V-PT dataset. We then pre-train the LMMs with subsets of the ShareGPT4V-PT and observe a clear positive relation between data volume and model performance.

Based on the above efforts and observations, we present ShareGPT4V-7B. It is pre-trained on the ShareGPT4V-PT and fine-tuned with ShareGPT4V and other datasets, resulting in superb modality alignment. Despite its simple model design, it outperforms 7B-scale competitors in all of the 11 benchmarks. In a nutshell, our contributions are threefold:

- We delve into LMM's training data and reveal three pivotal findings: 1) For data type, detailed captions are essential for efficient modality alignment. 2) For data quality, the detailed caption from cutting-edge LMMs or specially fine-tuned open-source LMMs can be close to humans and sufficient for efficient modality alignment. 3) For the data scale, the LMMs could have continuous improvement with the increased volume of detailed captions.
- We introduce the ShareGPT4V dataset, a large-scale image-text collection featuring 100K highly descriptive captions generated by GPT4-Vision, and the ShareGPT4V-PT dataset, consisting of 1.2M high-quality captions generated by our Share-Captioner. These datasets cast a light on the urgent need within the LMM community for high-quality captions to effectively align modalities.
- Leveraging the proposed dataset, we have developed the ShareGPT4V-7B, an advanced large multimodal model. Despite without elaborate architecture design, this model consistently demonstrates impressive performance across various multi-modal benchmarks.

2 Related Work

Large Language Models. In recent years, with the surge in data and computational power, the development of large language models has experienced a



COCO: Adults and children watching a train slowly leave.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure. There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.
A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.
The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

(a) Comparison of Captions' Quality

(b) Comparison of Performance

Fig. 1: (a) **We showcase a comparison between the caption** in our proposed ShareGPT4V dataset and those utilized by recent large multi-modal models (LMMs). Unlike COCO-Caption [9] involves brief human-made captions on the main subject. LLaVA-Instruct [33] combines human-made captions, bounding boxes, and GPT4 [41] to 'imagine' the image details, which leads to inevitable error/hallucination description (marked in red). Our approach involves feeding carefully designed prompts along with images directly into the advanced GPT4-Vision [42] and the descriptions are more detailed and accurate (marked in blue). (b) **We highlight the remarkable performance** of the proposed LMM, ShareGPT4V-7B, developed with the assistance of the ShareGPT4V dataset.

boom. Early encoder-decoder models like BERT [13] and T5 [49], and decoder-centric models such as GPT [48], leveraged the Transformer architecture [59] to excel in various NLP tasks. The success in GPT3 [4] has popularized the use of decoder-only architectures, which rely on auto-regressive decoding for generating predictions. Subsequent models like PaLM [11] extended the limits of model parameters and dataset scale, while others like InstructGPT [44] and ChatGPT [41] introduced fine-tuning and reinforcement learning techniques for improved conversational interaction. These developments, along with contributions from the open-source community [10, 56, 57, 57, 62], have set new benchmarks and opened avenues for future research in NLP area.

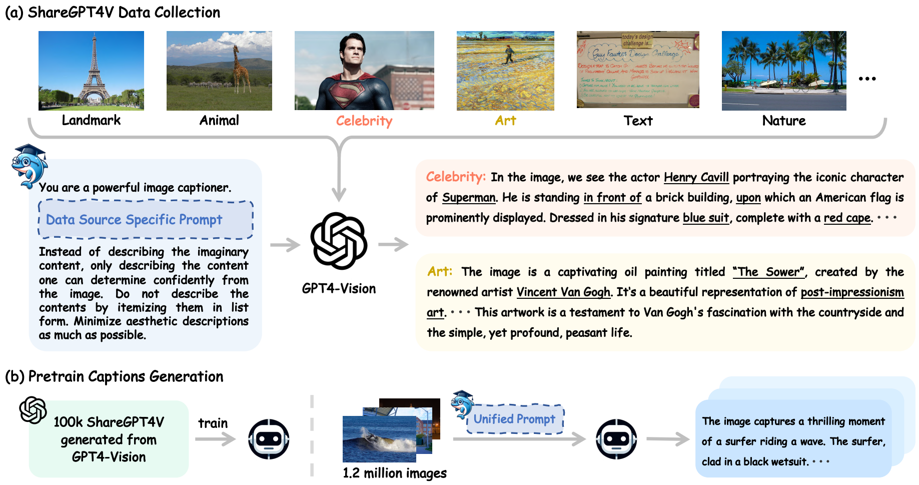
Large Multi-modal Models. As LLMs rapidly evolve, a faction within the research community is increasingly concentrating on introducing visual knowledge into LLMs. Central to this area are the seminal works in modality alignment within the vision-language learning area [21, 47]. A notable instance is CLIP [47], which exemplifies the alignment of visual and textual modalities through contrastive learning on extensive image-text pairs. A series of works [28, 29] were improved upon CLIP by employing refined data strategies for more diverse data, they have been effective for basic visual tasks [30, 34, 65] but less so for complex tasks like visual question answering. MiniGPT-4 [5], leveraging an LLM [10] and a visual encoder [16], has shown proficiency in image-text dialogues through pre-training alignment and instruction fine-tuning. Subsequent research [3, 6–8, 12, 26, 33, 45, 46, 63, 66] has further enhanced LMMs by focusing on the quality and diversity of pretraining and fine-tuning data. For instance, LLaVA [33] and InstructBLIP [12], with improved instruction fine-tuning, have advanced the understanding of complex prompts. mPLUG-Owl [63], Shikra [6], and KOSMOS-2 [45] have introduced new data types and training techniques, like grounding data, to reduce hallucinations and improve LMMs' grounding capability. Regrettably, it appears that the current LMMs have somewhat overlooked a crucial element: the quality of captions in image-text pairs.

Image-text Data Enhancement. In the vision-language learning area, several initiatives [15, 18, 25, 40] have been undertaken to enhance the quality of captions within image-text pairs. LaCLIP [15] leverages LLMs to rewrite raw captions, but its effectiveness is often hindered by hallucinations due to limited visual information and the low quality of original captions. Research [18, 40] explores methods to filter and blend raw and synthetic captions to enhance the CLIP model. A recent work, VeCLIP [25], proposes using LLMs to amalgamate information from both raw and synthetic captions. Nevertheless, the approach is constrained by the low quality of synthetic captions. DCI [58] curated 8K detailed captions on the SAM [23] dataset by human annotators, resulting in expansive costs. To the best of our knowledge, in the LMM area, LLaVA [33] uniquely inputs human-annotated short captions and bounding boxes into the GPT4 language model. This approach lets the model 'imagine' viewing the image before producing detailed captions. However, this method relies heavily on extensive human-annotated data and does not allow the model to truly 'see' the images. Consequently, it tends to generate detailed descriptions primarily of main objects, often including those in obscure corners but annotated with bounding boxes, leading to potential hallucinations in the LMMs' output. In contrast, we employ the most advanced LMM, GPT4-Vision, which is capable of directly producing highly descriptive captions from deliberated prompts and corresponding image inputs.

3 ShareGPT4V Series

3.1 Overview

In this section, we detail the process of developing the ShareGPT4V series. Subsection 3.2 elaborates how we created the **ShareGPT4V** dataset, including how we harnessed GPT4-Vision to generate 100K high-quality captions from various image sources. Subsection 3.3 describes our expansion from 100K high-quality captions to 1.2M to form the **ShareGPT4V-PT** dataset, matching the quality generated by GPT4-Vision with acceptable cost. Subsection 3.4 then explains how we leveraged the aforementioned datasets and tailored training strategies based on the characteristics of detailed captions to develop the **ShareGPT4V-7B** model, a simple yet superior LMM.


(a) ShareGPT4V Data Collection
Landmark
Animal
Celebrity
Art
Text
Nature
...
Data Source Specific Prompt
GPT4-Vision
(b) Pretrain Captions Generation
100k ShareGPT4V generated from GPT4-Vision
train
1.2 million images
Unified Prompt

You are a powerful image captioner.

Instead of describing the imaginary content, only describing the content one can determine confidently from the image. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible.

Celebrity: In the image, we see the actor Henry Cavill portraying the iconic character of Superman. He is standing in front of a brick building, upon which an American flag is prominently displayed. Dressed in his signature blue suit, complete with a red cape. · · ·

Art: The image is a captivating oil painting titled "The Sower", created by the renowned artist Vincent Van Gogh. It's a beautiful representation of post-impressionism art. · · · This artwork is a testament to Van Gogh's fascination with the countryside and the simple, yet profound, peasant life.

The image captures a thrilling moment of a surfer riding a wave. The surfer, clad in a black wetsuit. · · ·

Fig. 2: An overview for crafting the ShareGPT4V dataset. (a) We illustrate the procedure for collecting highly descriptive captions from GPT4-Vision [42] via various image sources and data-specific prompts, resulting in 100K high-quality captions that encapsulate a wide array of information conveyed by the images. (b) We delineate the process of utilizing the seed captions to train a general captioner and then employing this captioner to generate 1.2M high-quality captions for pre-training usage.

3.2 ShareGPT4V Dataset

The data collection pipeline of ShareGPT4V is shown in Figure 2(a). For each image selected from a specific data source D, we employed a meticulously crafted, data-specific prompt P_D. This prompt instructed GPT4-Vision to generate detailed descriptions, taking into account factors such as world knowledge, object attributes, spatial relationships, and aesthetic evaluations.

Data sources. To maximize the diversity and comprehensiveness of our data, we compiled around 100K images from various data sources, including images for detection [31] and segmentation [23], complex text-containing images [54], as well as various web images [43,51,53] containing artworks, landmarks, celebrities *etc*. More details can be found in the supplementary material.

Prompt Design. Given the diversity of our image sources, we expect a highly content-related description for each image. That is, the captions should extend beyond mere appearance and attributes, incorporating knowledge-related information. For instance, the Eiffel Tower should not be simply described as a tall iron tower, and a picture of Einstein should not be concluded as an old man.

For the description quality and stability, we designed a base prompt for a general description and added a specialized prompt for each data source. The base prompt asks the GPT4-Vision to describe the basic information of the image, including the object attributes, appearance, and spatial relationships. The specialized prompt focuses on some data-related information, as shown in Fig-

Table 2: Comparison of widely-used caption datasets and our proposed ShareGPT4V series datasets. 'LCS' abbreviates the LAION [51], CC [53], and SBU [50] datasets. The 'Visible' column denotes the image visibility during captioning, and the last column 'Average' shows the average character number of the caption.

Name	Image Source	Visible	Captioned by	Samples	Average
COCO-Caption [9]	COCO [31]	✓	Human	118K	52
BLIP2-LCS [29]	LCS	✓	BLIP2 [29]	558K	54
LLaVA-23K [33]	COCO [31]	×	GPT4 [41]	23K	609
ShareGPT4V	LCS, COCO [31], etc	✓	GPT4-Vision [42]	100K	**942**
ShareGPT4V-PT	LCS, COCO [31], etc	✓	Share-Captioner	**1,246K**	826

ure 2, we emphasize that the GPT4-Vision should mention some corresponding knowledge, such as the name and geographical location of a landmark-related image. Additionally, we add an aesthetic-related prompt for part of the images, to further improve the comprehensiveness of the description.

3.3 ShareGPT4V-PT Dataset

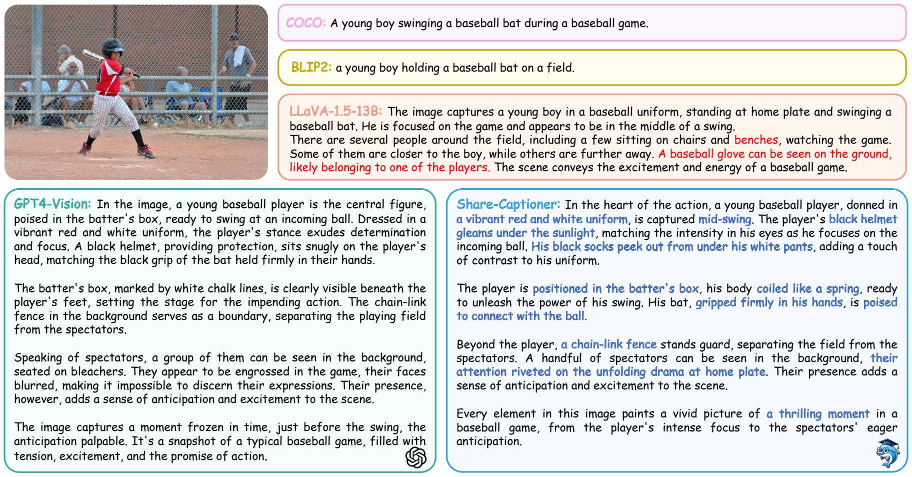
Compared with the supervised fine-tuning stage, modality alignment in the pre-training phase is more crucial and demands a large-scale dataset. For building a pre-training dataset, we employed the 100K high-quality captions generated by GPT4-Vision to fine-tune an alternative caption model and we have named it as Share-Captioner. Thanks to its training on diverse and comprehensive data, the Share-Captioner is capable of generating highly content-related descriptions with unified instruction. This approach allows the data scaling phase to proceed without the need for specialized prompt design.

To amass a substantial volume of high-quality image-text pairs, we selected a subset of 1.2 million images from current public datasets (see supplementary material for more details) and employed our pre-trained Share-Captioner for the captioning process. The entire caption generation process required around 44 A100 GPU days and we name this part of data as ShareGPT4V-PT.

Qualitative Analysis. For qualitative analysis, Figure 3 presents caption results from human-made COCO-Captions [9], BLIP2 [28], LLaVA-1.5-7B [32], Share-Captioner, and GPT4-Vision. It is important to note that the images featured in this figure were not part of the training dataset for Share-Captioner. The results depicted in Figure 3 demonstrate that Share-Captioner produced results that are closely comparable to those generated by GPT4-Vision, aligning with our anticipated capabilities for the captioning process.

3.4 ShareGPT4V-7B Model

To ascertain the efficacy of the curated datasets, we conducted experiments within a fair and controlled setting. This led to the development of ShareGPT4V-



COCO: A young boy swinging a baseball bat during a baseball game.

BLIP2: a young boy holding a baseball bat on a field.

LLaVA-1.5-13B: The image captures a young boy in a baseball uniform, standing at home plate and swinging a baseball bat. He is focused on the game and appears to be in the middle of a swing.
There are several people around the field, including a few sitting on chairs and benches, watching the game. Some of them are closer to the boy, while others are further away. A baseball glove can be seen on the ground, likely belonging to one of the players. The scene conveys the excitement and energy of a baseball game.

GPT4-Vision: In the image, a young baseball player is the central figure, poised in the batter's box, ready to swing at an incoming ball. Dressed in a vibrant red and white uniform, the player's stance exudes determination and focus. A black helmet, providing protection, sits snugly on the player's head, matching the black grip of the bat held firmly in their hands.

The batter's box, marked by white chalk lines, is clearly visible beneath the player's feet, setting the stage for the impending action. The chain-link fence in the background serves as a boundary, separating the playing field from the spectators.

Speaking of spectators, a group of them can be seen in the background, seated on bleachers. They appear to be engrossed in the game, their faces blurred, making it impossible to discern their expressions. Their presence, however, adds a sense of anticipation and excitement to the scene.

The image captures a moment frozen in time, just before the swing, the anticipation palpable. It's a snapshot of a typical baseball game, filled with tension, excitement, and the promise of action.

Share-Captioner: In the heart of the action, a young baseball player, donned in **a vibrant red and white uniform**, is captured **mid-swing**. The player's **black helmet gleams under the sunlight**, matching the intensity in his eyes as he focuses on the incoming ball. **His black socks peek out from under his white pants**, adding a touch of contrast to his uniform.

The player is **positioned in the batter's box**, his body **coiled like a spring**, ready to unleash the power of his swing. His bat, **gripped firmly in his hands**, is **poised to connect with the ball**.

Beyond the player, **a chain-link fence** stands guard, separating the field from the spectators. A handful of spectators can be seen in the background, **their attention riveted on the unfolding drama at home plate**. Their presence adds a sense of anticipation and excitement to the scene.

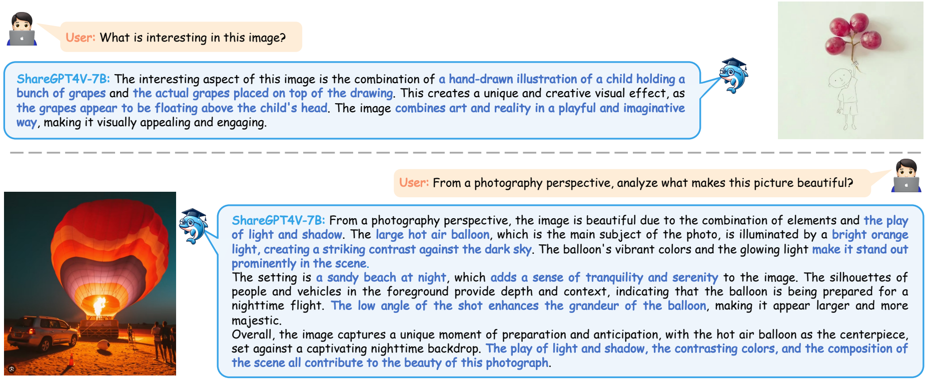
Every element in this image paints a vivid picture of **a thrilling moment** in a baseball game, from the player's intense focus to the spectators' eager anticipation.

Fig. 3: A qualitative comparison of caption quality from various sources. The COCO [9] captions were generated by humans and the remaining captions were produced by large multi-modal models using the same prompts and images. Mistakes within the captions are highlighted in **red**, whereas detailed and accurate parts are emphasized in **blue**. Notably, the image used for this comparison was not included in the training set of our Share-Captioner.

7B, a streamlined yet superior baseline LMM leveraging our high-quality data in both the pre-training and SFT stages.

Model Architecture. The ShareGPT4V-7B model follows the design of LLaVA-1.5 [32], including three integral components: (1) A vision encoder utilizing the CLIP-Large model [47], with a resolution of 336×336 and a patch size of 14, converting input images into 576 tokens. (2) A projector, which is a two-layer multi-layer perception (MLP), is introduced to connect the vision and language modalities. (3) A LLM, based on the open-source Vicuna-v1.5 [10], derived from LLaMA2 [57]. Currently, we focus on the lightweight 7B model, and we have empirically validated that even with lightweight training data and model scale, it can significantly outperform many current LMMs that utilize extensive training datasets or larger model scales.

Pre-Training. In the pre-training stage, we utilize the pre-training subset of the ShareGPT4V dataset, *i.e.* ShareGPT4V-PT. Given these high-quality captions, solely fine-tuning the MLP does not suffice to exploit their full capabilities. In previous LMM research [5, 32, 33, 69], the vision encoder is generally not fine-tuned during pre-training, a rational approach considering the lower quality of previously used captions, where fine-tuned the vision encoder might degrade its visual knowledge extraction ability. We opted for simultaneous fine-tuning of the vision encoder, projector, and large language model. With this configuration, the large language model acquires a native understanding of visual embeddings, while also prompting the vision encoder to create relevant visual embeddings



User: What is interesting in this image?

ShareGPT4V-7B: The interesting aspect of this image is the combination of a hand-drawn illustration of a child holding a bunch of grapes and the actual grapes placed on top of the drawing. This creates a unique and creative visual effect, as the grapes appear to be floating above the child's head. The image combines art and reality in a playful and imaginative way, making it visually appealing and engaging.

User: From a photography perspective, analyze what makes this picture beautiful?

ShareGPT4V-7B: From a photography perspective, the image is beautiful due to the combination of elements and the play of light and shadow. The large hot air balloon, which is the main subject of the photo, is illuminated by a bright orange light, creating a striking contrast against the dark sky. The balloon's vibrant colors and the glowing light make it stand out prominently in the scene.

The setting is a sandy beach at night, which adds a sense of tranquility and serenity to the image. The silhouettes of people and vehicles in the foreground provide depth and context, indicating that the balloon is being prepared for a nighttime flight. The low angle of the shot enhances the grandeur of the balloon, making it appear larger and more majestic.

Overall, the image captures a unique moment of preparation and anticipation, with the hot air balloon as the centerpiece, set against a captivating nighttime backdrop. The play of light and shadow, the contrasting colors, and the composition of the scene all contribute to the beauty of this photograph.

Fig. 4: Examples of multi-modal dialogue with ShareGPT4V-7B model. ShareGPT4V-7B exhibits satisfactory capabilities in understanding image details and performing aesthetic assessments. High-quality content is highlighted in **blue**.

for elements in captions. This setup enables a comprehensive exploration and understanding of the knowledge embedded in visual embeddings, aligned with the intricate details of the captions. Specifically, we consistently applied a learning rate of $2e^{-5}$ across all components, with a batch size set at 256, and the comprehensive optimization process spanned roughly 4700 steps. Notably, we experimentally found that selectively fine-tuning only the latter half of the vision encoder's layers achieves optimal results, coupled with a satisfactory level of training efficiency.

Supervised Fine-Tuning. As we emphasized above, the goal of this paper is not to build a new SOTA model with some unique architecture designs but to **explore the influence of training data on realizing efficient modality alignment.** So we utilize the 665K supervised data organized by LLaVA-1.5 and only replace part of it with our ShareGPT4V dataset. In detail, the 665K data is gathered from publicly available academic task-oriented data [1,22,24,38, 39,52,54] and instruction-tuning data for conversational and complex reasoning tasks [33] involving natural images [31]. It contains 23K detailed description data and we replaced it with randomly sampled 23K high-quality captions from the 100K captions in ShareGPT4V. During the SFT stage, to enhance the training efficiency and compare fairly, we froze the vision encoder and instead focused on fine-tuning the projector and the large language model. The learning rate was established at $2e^{-5}$, with a batch size of 128, and the total optimization process spanned around 5200 steps.

4 Experiments

4.1 Benchmarks

To thoroughly assess our proposed ShareGPT4V-7B model, we evaluate it across 11 benchmarks, covering a range of academic Visual Question Answering (VQA)

Table 3: Comparison with SoTA methods on 11 benchmarks. With 7B parameters, ShareGPT4V-7B outperforms competitors in 9 out of 11 benchmarks and ranks second on the others, despite these competitors using larger training datasets or more parameters. Benchmark names are abbreviated due to space limits. LLaVA^W: LLaVA-Bench (In-the-Wild) [33]; MME^P: MME Perception [17]; MME^C: MME Cognition [17]; MMB: MMBenchmark [35]; MMB^{CN}: MMBench-Chinese [35]; SEED^I: SEED-Bench (Image) [27]; MM-Vet [64]; QBench [61]; SQA^I: ScienceQA-IMG [36]; VQA^{V2} [19]; VizWiz [20]. * indicates our re-implemented test results missed in benchmarks or origin papers. The best results are **bold** and the second-best results are underlined.

Method	Language Model	LLaVA^W	MME^P	MME^C	MMB	MMB^{CN}	SEED^I	MM-Vet	QBench	SQA^I	VQA^{V2}	VizWiz
BLIP-2	FLAN-T5	38.1	1293.8	290.0	-	-	46.4	22.4	-	61.0	41.0	19.6
InstructBLIP	Vicuna-7B	60.9	-	-	36.0	23.7	53.4	26.2	56.7	60.5	-	34.5
InstructBLIP	FLAN-T5	58.2	1212.8	291.8	-	-	-	25.6	-	63.1	-	33.4
Shikra	Vicuna-13B	-	-	-	58.8	-	-	-	54.7	-	77.4	-
IDEFICS-80B	LLaMA-65B	-	-	-	54.5	38.1	-	-	-	-	60.0	36.0
Qwen-VL	Qwen-7B	-	-	-	38.2	7.4	56.3	-	59.4	67.1	78.8	35.2
Qwen-VL-Chat	Qwen-7B	-	1487.5	360.7	60.6	56.7	58.2	-	-	68.2	78.2	38.9
mPLUG-Owl2	LLaMA2-7B	59.9*	1450.2	336.2*	64.5	60.3*	64.5*	36.2	62.9	-	79.4	54.5
LLaVA	Vicuna-7B	63.0*	807.0*	247.9*	34.1*	14.1*	25.5*	26.7*	-	38.5*	79.0*	9.3*
LLaVA-1.5	Vicuna-7B	63.4	1510.7	316.1*	64.3	58.3	66.2*	30.5	58.7	66.8	78.5	50.0
LLaVA-1.5	Vicuna-13B	70.7	1531.3	295.4*	67.7	**63.6**	68.2	35.4	62.1	**71.6**	80.0	53.6
ShareGPT4V-7B	Vicuna-7B	**72.6**	**1567.4**	**376.4**	**68.8**	62.2	**69.7**	**37.6**	**63.4**	68.4	**80.6**	**57.2**

tasks and recent benchmarks designed specifically for large multi-modal models (LMMs). The LLaVA (in the wild) benchmark [33] is composed of 60 questions, spanning three distinct tasks: conversation, complex reasoning, and detailed description. The MME Benchmark [17] evaluates LMMs' perception and cognition capabilities through a series of carefully crafted questions across 14 sub-tasks. MMBench and MMBench-CN [35] benchmarks manually design questions to evaluate the model's vision-related reasoning and perception abilities for English and Chinese, respectively. SEED [27], with the assistance of GPT4, generated a dataset comprising approximately 19K questions related to images and videos. MM-Vet [64] uses GPT4 for a six-dimensional LMM capability assessment. Q-Bench [61] assesses low-level perception, while VQA-v2 [19] and VisWiz [20] are benchmarks in the realm of traditional Visual Question Answering (VQA) tasks.

4.2 Quantitative Comparison of LMMs

As illustrated in Table 3, we present a quantitative comparison between our proposed ShareGPT4V-7B model and existing state-of-the-art LMMs. Notably, compared with previous LMMs, our ShareGPT4V-7B attained the most superior performance in 9 out of the total 11 benchmarks.

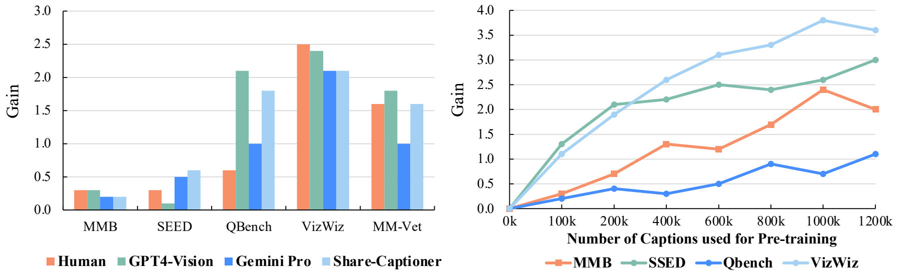
Specifically, our ShareGPT4V-7B model outperformed the previously best-performing LLaVA-1.5-13B model by 1.9 points on the LLaVA (in the wild) benchmark, demonstrating superior capabilities in tasks such as detailed description and complex reasoning. On the MME Benchmark, it achieved the highest scores in both perception (P) and cognition (C) capabilities, surpassing LLaVA-1.5-13B in perception by 36.1 points and exceeding Qwen-VL-Chat, which was trained on 1.4 billion data, by 15.7 points in cognition. Our model also achieved

Table 4: Comparison of lexical composition of the captions generated by GPT4-Vision and Share-Captioner.

Lexical Category	n.	adj.	adv.	v.	num.	prep.
GPT4-Vision	27.3%	9.5%	2.0%	12.3%	0.5%	11.4%
Share-Captioner	27.4%	8.8%	1.5%	12.5%	0.4%	11.5%

Table 5: Human preference on Share-Captioner vs. GPT4-Vision over 100 validation samples and 10 volunteers.

	GPT4-Vision	Share-Captioner	Comparable
Percentage	38.2%	35.3%	26.5%
Avg. Score	2.2	1.8	-



(a) Comparison of caption quality from various sources (b) Performance gains from increasing number of captions

Fig. 5: (a) **Comparison of caption sources.** LLaVA-1.5-7B serves as the baseline and we add the same images with various captions into the training data to investigate the performance gap between LMMs and human annotators. (b) **Scaling curve of performance gains corresponding to pre-training data volume.** The model shows consistent gain with the number of high-quality captions scales up.

an optimal accuracy of 68.8% on MMBench, leading the second-best by 1.1%. Furthermore, on the SEED (image) benchmark, which includes 9 assessment dimensions and 14K questions, ShareGPT4V-7B achieved the highest score of 69.7%, 1.5% higher than the second-ranked LLaVA-1.5-13B. In the low-level image assessment QBench, our model's top score of 63.4% can be attributed to the diversity of our constructed dataset. Lastly, our model almost consistently performed best on traditional VQA benchmarks with the smallest model size.

Our findings demonstrate that even with a simple architecture, public data, and lighter parameters (7B), it is possible to outperform many competitors with massive training data and parameter sizes, thanks to the support of these high-quality captions.

5 Discussions

5.1 Caption quality analysis

Can LMMs be effective alternatives for humans on the detailed-caption task? To investigate the performance gap between professional human annotators and LMM alternatives, we collect a series of 8K captions of the SAM [23] image dataset from human [58], GPT4-Vision [42], Gemini Pro [55], and our Share-Captioner. We then integrate these captions into the SFT data of baseline (LLaVA-1.5-7B). Surprisingly, as shown in Figure 5(a), existing state-of-the-art

Table 6: Effectiveness of ShareGPT4V Series Datasets in each stage. The ShareGPT4V dataset improves the model performance in both the pre-training and supervised fine-tuning stages.

Pre-training with ShareGPT4V-PT	SFT with ShareGPT4V	MME^P	MMBench	SEED^I
✗	✗	1510.7	64.3	66.2
✗	✓	1542.1	66.8	66.7
✓	✗	1557.2	67.4	68.5
✓	✓	**1567.4**	**68.8**	**69.7**

LMMs can be close to human annotators' capabilities. Furthermore, by applying lightweight fine-tuning to open-source LMMs, we can obtain a caption model that can get close to human performance, thereby alleviating both labor and economic costs.

Qualitative captioning capability comparison between Share-Captioner and GPT4-Vision. We first analyze the lexical composition of the captions produced by GPT4-Vision and Share-Captioner, and the results are presented in Table 4. The analysis reveals that the captions generated by our Share-Captioner contain a comparable information level to those generated by GPT4-Vision. Furthermore, as shown in Table 5, we first generate 100 captions with GPT4-Vision and Share-Captioner and then invite 10 volunteers to evaluate the captions based on three aspects. These aspects include: (1) **Omission** - checking for no key elements missing in the caption; (2) **Fabrication** - identifying imaged elements not present in the image; and (3) **Distortion** - assessing the accuracy of element attributes such as color and size. Each pair can earn a maximum of 3 points, one for each criterion met. As anticipated, our Share-Captioner performs on par with the GPT4-Vision, confirming the quality of the ShareGPT4V-PT dataset.

5.2 ShareGPT4V dataset analysis

Effectiveness of ShareGPT4V Series Datasets. As shown in Table 6, we conducted a thorough ablation study to assess the impact of the ShareGPT4V-PT and ShareGPT4V subsets. Our baseline is the LLaVA-1.5-7B model, without utilizing the ShareGPT4V dataset in either pretraining or SFT stages. Utilizing only our ShareGPT4V subset during the SFT stages resulted in a significant increase of 31.4 points in MME perception score, and improvements of 2.5% and 0.5% in accuracy on the MMBench and SEED benchmarks, respectively. Notably, ShareGPT4V used here was selected from various data sources, yielding more performance gains than those from solely the COCO dataset (see in Figure 6). When only the ShareGPT4V-PT subset was used during pretraining, we observed a remarkable gain of 46.5 points in MME perception, along with substantial accuracy improvements of 3.1% and 2.3% on the MMBench and SEED benchmarks, respectively. Moreover, employing the ShareGPT4V dataset in both pretraining and SFT phases led to further satisfactory enhancements in overall

Table 7: Ablation on the pre-training caption quality. Based on the baseline, the second and third rows share the same end-to-end training strategy and images, but different captions from the BLIP2 captioner or our ShareGPT4V-PT dataset.

Method	MME^P	MMBench	SEED^I
Basline	1516.9	65.3	66.8
+BLIP2-558K	1521.6	66.2	66.9
+ShareGPT4V-PT-558K	**1539.8**	**68.3**	**68.9**

performance, effectively validating the necessity of incorporating high-quality captions in both training stages.

Influence of Caption Scale in Pre-training. In Figure 5(b), we present our investigation into the required quantity of high-quality captions for the pre-training stage. Here we randomly sample the data from the ShareGPT4V-PT and train the model with the subset, which varies from 100K to 1200K. The results show that with only 100K high-quality data, the model has a significant improvement on all the benchmarks, this further proves the effectiveness of the high-quality data. When the model is pre-trained with more than 1000K data, its performance improvement slows down while it keeps increasing in part of the benchmarks. This indicates the further scaling potential of the model and data, and we left it for further exploration.

Influence of Caption Quality in Pre-training. Then we study how the caption quality influences the pre-training performance. For a fair comparison, we pre-train the model with the same setting and images, but the captions are generated by different models. In detail, we use the 558K LAION-CC-SUB image-text pairs captioned by the BLIP2 as the baseline and replace the text with the high-quality one in our ShareGPT4V-PT. As results shown in Table 7, comparing with the baseline, the joint training strategy with the BLIP2-558K data gets better results on all the benchmarks, while the gain is quite minor that only 4.7 in MME Perception and 0.1 on SEED Bench. When we replace the captions with our ShareGPT4V-PT-558K, the model gets significant gains. In detail, it gets 1549.8, 68.3, 68.9 on the three benchmarks, surpassing the BLIP2-558K case with 18.2, 1.9 and 2.0 respectively. This proves the essential of high-quality captions for effective pre-training and modality alignment.

Influence of LLMs Architecture. Then we study the gain of ShareGPT4V among different LLMs. we chose various advanced, publicly available LMMs with various architectures, including LLaVA-7B [33], LLaVA-1.5-7B [32], LLaVA-1.5-13B [32], and Qwen-VL-Chat-7B [3]. For a fair comparison, we equally replaced the detailed captions in their SFT datasets with a fixed selection from our 100K GPT4-Vision-generated captions, while maintaining image sources as consistent as possible. As depicted in Figure 6, the integration of our highly descriptive captions consistently improves the SFT phase performance across these varied LMMs. This further proves that the gain from high-quality data is universal.

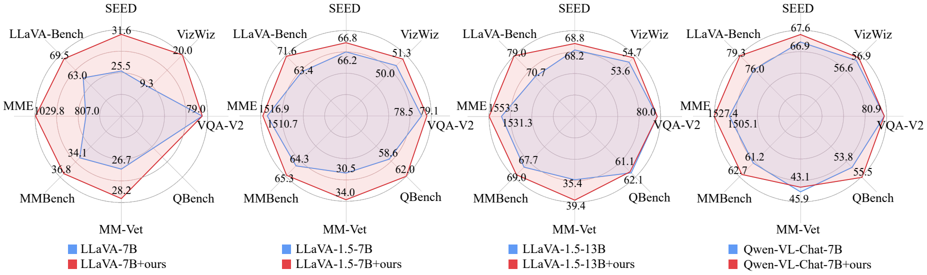


Fig. 6: The gain from high-quality captions is universal among model architecture. We compare the performance of various large multi-modal models before and after equally replacing their SFT captions with a subset of our ShareGPT4V.

Table 8: Influence of learnable ViT block number. The LLM gets better results with half of the blocks learnable during the pre-training.

Tune from Block	Memory Usage	MME^P	MMBench	SEED^I
24	49.6 GB	1515.2	66.6	68.1
18	53.2 GB	1556.0	67.2	69.3
12	56.7 GB	**1567.4**	**68.8**	**69.7**
6	60.0 GB	1529.5	67.7	69.6
0	63.6 GB	1545.7	68.5	69.2

Influence of Learnable ViT Blocks Number. As detailed in Table 8, we extensively investigated the optimal approach for fine-tuning the vision encoder during the pre-training stage. Compared to freezing the vision encoder during the pre-training, we found that for high-quality captions, unlocking the vision encoder facilitates more effective modality alignment.

6 Conclusion

In this paper, we explore the impact of training data attributes on LMMs, unveiling three critical insights. First, we find that detailed captions significantly boost LMMs' understanding and reasoning capabilities. Second, lightweight fine-tuning with high-quality captions brings open-source LMM to be a powerful captioner close to humans. Third, leveraging the local captioner to increase the volume of detailed captions consistently enhances LMM performance. Associated with these investigations, we created the ShareGPT4V series, comprising a 100K high-quality caption dataset from GPT4-Vision, a powerful local caption model, a 1.2M high-quality caption dataset produced by local caption model, and a simple yet superior LMM built on these resources. We are committed to making ShareGPT4V fully accessible to the public, with the aspiration that it becomes a foundational resource in advancing the field of LMMs.

Acknowledgements

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. This work is also partially supported by the National Key R&D Program of China (2022ZD0160201), Shanghai Artificial Intelligence Laboratory, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

1. Sharegpt. `https://sharegpt.com/` (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Chen, J., Li, D.Z.X.S.X., Zhang, Z.L.P., Xiong, R.K.V.C.Y., Elhoseiny, M.: Minigpt-v2: Large language model as a unified interface for vision-language multitask learning. arXiv preprint arXiv:2310.09478 (2023)
6. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
7. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330 (2024)
8. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., et al.: Sharegpt4video: Improving video understanding and generation with better captions. arXiv preprint arXiv:2406.04325 (2024)
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
10. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
11. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
12. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

14. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360 (2021)
15. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. arXiv preprint arXiv:2305.20088 (2023)
16. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)
17. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
18. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
19. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
20. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
22. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
25. Lai, Z., Zhang, H., Wu, W., Bai, H., Timofeev, A., Du, X., Gan, Z., Shan, J., Chuah, C.N., Yang, Y., et al.: From scarcity to efficiency: Improving clip training via visual-enriched captions. arXiv preprint arXiv:2310.07699 (2023)
26. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
27. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

30. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
35. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
36. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
37. Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models. arXiv preprint arXiv:2305.15023 (2023)
38. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
39. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019)
40. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning. arXiv preprint arXiv:2307.10350 (2023)
41. OpenAI: Chatgpt. `https://chat.openai.com/` (2023)
42. OpenAI: Gpt-4v(ision) system card (2023)
43. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011)
44. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
45. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
46. Qiao, Y., Duan, H., Fang, X., Yang, J., Chen, L., Zhang, S., Wang, J., Lin, D., Chen, K.: Prism: A framework for decoupling and assessing the capabilities of vlms. arXiv preprint arXiv:2406.14544 (2024)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

48. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
49. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
50. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)
51. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
52. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
53. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
54. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 742–758. Springer (2020)
55. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
56. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities (2023)
57. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
58. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. arXiv preprint arXiv:2312.08578 (2023)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
60. Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., Jiang, Y.G.: To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574 (2023)
61. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181 (2023)
62. Yang, A., Xiao, B., Wang, B., Zhang, B., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., et al.: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
63. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)

64. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
65. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems **35**, 36067–36080 (2022)
66. Zhang, P., Dong, X., Zang, Y., Cao, Y., Qian, R., Chen, L., Guo, Q., Duan, H., Wang, B., Ouyang, L., et al.: Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. arXiv preprint arXiv:2407.03320 (2024)
67. Zhang, P., Wang, X.D.B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., et al.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 (2023)
68. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
69. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)