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)

A Data Sources

Data Source Composition for ShareGPT4V. To maximize the comprehensiveness of our captions, we compiled a total of 100K images from diverse sources. This includes 50K images from COCO [2], 30K images from 'LCS' (which abbreviates LAION [6], CC-3M [7], and SBU [4]), 20K images from SAM [1], 500 images from TextCaps [8], 500 images from WikiArt [5], and 1K images from web-crawled data (split evenly between images of landmarks and images of celebrities).

Data Source Composition for ShareGPT4V-PT. We utilized our pre-trained Share-Captioner to generate the pre-training dataset. This dataset is comprised of a subset of 1.2M images selected from existing public datasets. These include 118K images from COCO [2],570K images from SAM [1], and 558K images from LLaVA-1.5 pre-training data [3].

B Extra experiments

How about using the full ShareGPT4V dataset in SFT? We experimented with replacing the original 23K captions with the entire 100K detailed captions in SFT, resulting in 100K/742K detail-caption/training data. As shown in Table 1, more high-quality data can further improve the model performance on all the tasks. The experiments conducted in the main paper primarily aimed at ensuring a fair comparison with baseline methods. Therefore we decided to use an equivalent amount of 23K high-quality captions for substitution. Future researchers could utilize the full 100K ShareGPT4V dataset to further enhance the performance of their LMMs.

Is it the more detailed captions or fewer hallucinations that lead to better modality alignment? To align with LLaVA-1.5's detail level roughly, we truncated 23K ShareGPT4V captions by LLaVA-1.5's average caption length. We posited that such truncation holds a fixed hallucination rate. As shown in Table 2, both factors aid modality alignment and the detailness slightly bringing more gain.

* Equal contribution † Corresponding authors.

Table 1: Comparison of replacing with 23K and 100K captions.

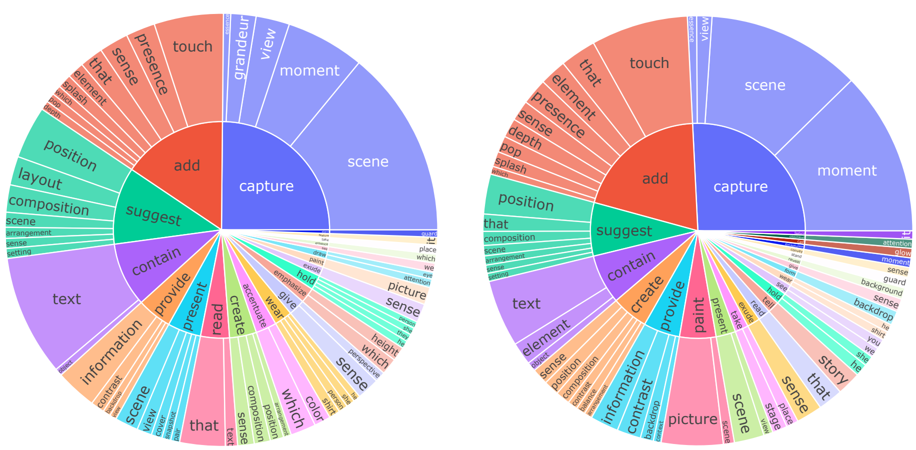
Method	detailed caption in SFT data	MME^P	MMB	SEED	MM-Vet	LLaVA^W
LLaVA-1.5-7B	Original 23K	1510.7	64.3	66.2	30.5	63.4
LLaVA-1.5-7B	ShareGPT4V 23K	1516.9	65.3	66.8	34.0	71.6
LLaVA-1.5-7B	ShareGPT4V 100K	**1540.3**	**65.9**	**68.8**	**36.7**	**72.3**

Table 2: Compare detailness and hallucination. ‘tr’ for truncated.

Method	detailed caption in SFT data	MME^P	MMB	SEED	MM-Vet	LLaVA^W
LLaVA-1.5-7B	Original 23K	1510.7	64.3	66.2	30.5	63.4
LLaVA-1.5-7B	ShareGPT4V-tr 23K	1513.1	64.6	66.4	32.3	67.5
LLaVA-1.5-7B	ShareGPT4V 23K	**1516.9**	**65.3**	**66.8**	**34.0**	**71.6**

C Caption Analysis

Figure 1 provides a visualization of the root noun-verb pairs for the captions generated by both GPT4-Vision and Share-Captioner. It’s clear to see that the diversity and linguistic expression of the captions produced by Share-Captioner are comparable to those of GPT4-Vision.



(a) GPT4-Vision Generated Caption (b) Share-Captioner Generated Caption

Fig. 1: Analysis of captions generated by GPT4-Vision and Share-Captioner. Visualization of the root noun-verb pairs (occurring over 1%) of the captions.

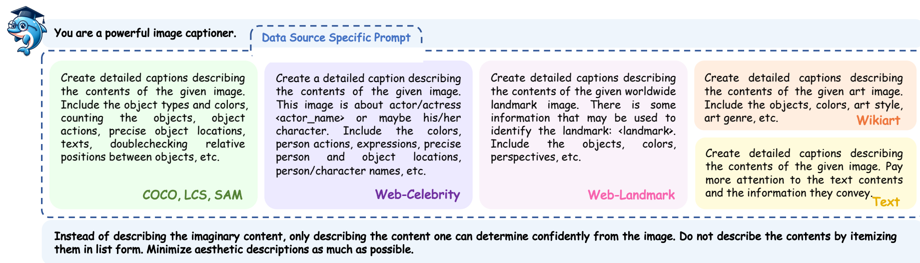


Fig. 2: Prompts for instructing GPT4-Vision to generate detailed descriptions. The Prompts are designed with base prompts at the beginning and end, with a data-specific prompt placed in between.

D Prompts

Given the diversity of our image sources, we expect a highly content-related description for each image. As shown in Figure 2, we designed a base prompt for a general description and added a specialized prompt for each data source.

E Details about Share-Captioner

Our captioner is built on the open-source LMM, InternLM-XComposer [9]. It contains an LLM InternLM-7B, a ViT EVG-g 1B, and a small Perceive Sampler. Specifically, we increased its resolution from 224×224 to 448×448 by fine-tuning the vision encoder to match the detail required for high-quality captions. It was trained on the 100K ShareGPT4V dataset with a learning rate of 2e-5, a batch size of 1024, over 2 epochs using 16 A100 GPUs for 3 hours. During training, we set a unified prompt for all captions, "`Analyze the image in a comprehensive and detailed manner`", to minimize the need for the data-specific prompt design.

F Examples

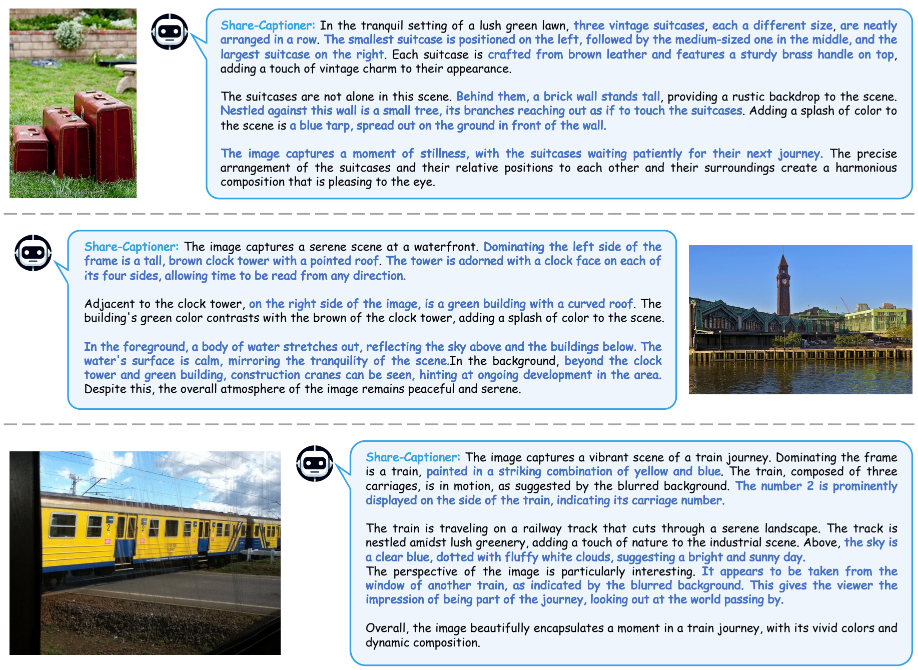


Fig. 3: Examples of captions generated by Share-Captioner. Detailed and accurate parts within the captions are emphasized in blue.

ShareGPT4V: 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. The painting vividly portrays a farmer in the act of sowing seeds in a field. The field, awash with hues of yellow and blue, forms a striking contrast against the farmer's figure. In the distance, a quaint village landscape unfolds, complete with houses and trees. The colors used are vibrant, bringing the scene to life with their intensity. The brushstrokes are clearly visible, adding texture and depth to the painting.

This artwork is a testament to Van Gogh's fascination with the countryside and the simple, yet profound, peasant life.

ShareGPT4V: The image captures a charming Halloween scene. Dominating the frame is a large, glowing orange pumpkin. It's been skillfully carved to display a classic jack-o'-lantern face, with triangular eyes and a toothy grin. The word "FERGUSON" is also carved into the pumpkin, perhaps indicating the name of the family who owns this decoration. The pumpkin's interior is illuminated, casting a warm light that accentuates the carved features.

Beside the orange pumpkin, there's a smaller white pumpkin. It's uncarved, providing a simple yet effective contrast to its larger, more elaborate neighbor. Adding a touch of whimsy to the scene is a silver owl figurine. Its metallic sheen stands out against the natural textures of the pumpkins.

All three objects are arranged on a wooden surface, which provides a rustic backdrop to the scene. Behind them, a gray wall further emphasizes the vibrant orange of the carved pumpkin. The relative positions of the objects suggest a thoughtful arrangement, with the glowing orange pumpkin as the centerpiece, flanked by the white pumpkin and the silver owl.

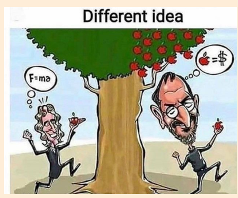
Despite the simplicity of the scene, it's clear that each object has been chosen and positioned with care, creating a Halloween decoration that's both festive and welcoming.

ShareGPT4V: The image captures a delightful scene of a meal, likely at a Chinese restaurant. The main focus is a white plate with a blue rim, presenting a dish of beef and broccoli. The beef, brown in color, and the broccoli, a vibrant green, are generously coated in a brown sauce, adding a glossy finish to the dish.

To the left of the plate, a spoon rests, ready to serve the delicious food. The plate itself is placed on a white tablecloth, providing a clean and simple backdrop that allows the colors of the food to stand out.

In the background, there are hints of other dishes, one of which appears to have a red sauce. These dishes add depth to the scene, suggesting a variety of flavors waiting to be explored. The image as a whole paints a picture of a hearty and appetizing meal, inviting the viewer to partake in the culinary experience.

Fig. 4: Examples of captions generated by ShareGPT4V. Detailed and accurate parts within the captions are emphasized in blue.

User: 

User: Who are the 2 characters in the picture?

ShareGPT4V-7B: The two characters in the picture are Steve Jobs and Isaac Newton.

User: What different ideas did they get from the apple?

ShareGPT4V-7B: Steve Jobs and Isaac Newton got the idea of the Apple computer and the concept of gravity, respectively, from an apple.

User: 

User: What is the most common catchphrase of the character on the right?

ShareGPT4V-7B: The most common catchphrase of the character on the right is "I am the one who knocks." This phrase is associated with the character Walter White from the TV show Breaking Bad, played by Bryan Cranston. Walter White is a methamphetamine manufacturer and distributor who is known for his intelligence, wit, and unpredictable behavior. The catchphrase is often used by Walter White to express his confidence and assertiveness, and it has become iconic in the series.

User: What kind of life would he have lived if he and Jesse Pinkman hadn't met?

ShareGPT4V-7B: If Walter White and Jesse Pinkman had never met, Walter White's life would have likely been very different. In the original timeline of the show, Walter White was a high school chemistry teacher who was diagnosed with cancer and decided to produce and sell methamphetamine to provide for his family after his death. He chose Jesse Pinkman, a former student and small-time criminal, as his partner in this venture.

Without Jesse, Walter would have had to find another partner or work alone to produce and distribute the methamphetamine. He might have had to rely on his own skills and knowledge to create the product and find customers. However, without Jesse's skills, connections, and resourcefulness, Walter's operation would likely have been smaller and less successful.

Additionally, without Jesse, Walter's character development would have been significantly altered. Jesse's presence and influence on Walter's life helped him to grow and change into the complex and morally ambiguous character that we see in the show. The relationship between Walter and Jesse is a central theme of Breaking Bad, and their dynamic drives much of the plot and character development.

Fig. 5: Example of Multi-Round Dialog of ShareGPT4V-7B. Detailed and accurate parts within the captions are emphasized in blue.

References

1. 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)
2. 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)
3. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
4. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)