SVIT: Scaling up Visual Instruction Tuning

Bo Zhao[*][†] **Boya Wu**[*] **Tiejun Huang**
Beijing Academy of Artificial Intelligence
{zhaobo, bywu}@baai.ac.cn
https://github.com/BAAI-DCAI/Visual-Instruction-Tuning

Abstract

Thanks to the emerging of foundation models, the large language and vision models are integrated to acquire the multimodal ability of visual captioning, dialogue, question answering, etc. Although existing multimodal models present impressive performance of visual understanding and reasoning, their limits are still largely under-explored due to the scarcity of high-quality instruction tuning data. To push the limits of multimodal capability, we Sale up Visual Instruction Tuning (SVIT) by constructing a dataset of 3.2 million visual instruction tuning data including 1.6M conversation question-answer (QA) pairs and 1.6M complex reasoning QA pairs and 106K detailed image descriptions. Besides the volume, the proposed dataset is also featured by the high quality and rich diversity, which is generated by prompting GPT-4 with the abundant manual annotations of images. We empirically verify that training multimodal models on SVIT can significantly improve the multimodal performance in terms of visual perception, reasoning and planing.

1 Introduction

The great success of large language models (LLMs), e.g. BERT[7], T5[23], GPT2[22], GPT3[3], have motivated the advancement of vision [8, 19, 11] and multimodal [21, 1, 30, 17] in terms of architecture design and learning paradigm. Recently, GPT-4 demonstrates impressive multimodal understanding and reasoning abilities, accepting image and text inputs while responding in text. Inspired by GPT-4, multimodal models bridging language and vision models have achieved remarkable progress in multiple visual understanding and reasoning tasks, e.g. visual captioning[15], dialogue[1] and question answering[30, 17].

Typically, the multimodal models are pre-trained on large multimodal datasets, e.g. LAION-2B[24], CC-12M[4], YFCC-100M[25] and mmc4[31], that contain millions to billions roughly-aligned image-text pairs from the web. Then, precise vision-language data pairs are used to finetune the models. Like the success of language instruction tuning, visual instruction tuning has become the key to the multimodal performance. However, due to the high construction cost, existing visual instruction datasets are still in small scale and less informative. Several works convert the image captioning and VQA datasets[16, 2, 12] into instruction tuning data by manually adding a few instructions [6]. However, these captions and questions/answers are usually short and focus on visual perception and simple questions, which may lead to ineffective model training[9]. To generate more informative visual instruction data, GPT models are introduced. LLaVA[17] contributes a large visual instruction dataset containing 150K data by prompting GPT-4 with five captions and a few object bounding boxes associated with images from COCO dataset[16]. Meanwhile, MiniGPT-4[30] creates 3,500 image-text pairs by refining model's output using ChatGPT. The language only GPT models have difficulty in precisely imagining the whole picture from the limited input. Thus, the generated instruction tuning data lacks diversity and complexity.

---

[*]Equal Contribution
[†]Corresponding Author

Preprint.

To push the limits of large multimodal models, we Scale up Visual Instruction Tuning (SVIT) and propose a large-scale dataset with 3.2 million informative instruction tuning data, including 1.6M conversation QA pairs, 1.6M complex reasoning QA pairs and 106K detailed descriptions. Tab. 1 shows that SVIT is 20× larger than LLaVA dataset. To enrich the diversity and informativeness of instruction tuning data, we construct SVIT based on Visual Genome which has abundant manual annotations and GPT-4 which has the best multimodal capability. We prompt the language-only GPT-4 ChatBot with image-level descriptions, detailed region descriptions and object bounding boxes. Extensive experiments verify that multimodal models tuned on the proposed SVIT dataset evoke impressive ability in visual perception, reasoning and planing.

Table 1: Comparing SVIT to similar vision-language instruction datasets.

Dataset	#Image	#Object BBox	#Region Description	#Image Caption	#Instruction Question	#Response Answer	GPT
MiniGPT-4	3.5K	-	-	-	4	3.5K	GPT-3.5
LLaVA	81.5K	600K	-	404.7K	150K	150K	GPT-4
SVIT	108.1K	3.8M	5.4M	257.6K	3.2M	3.2M	GPT-4

2 Related Work

2.1 Multimodal Models

Existing multimodal solutions can be roughly split into two categories: 1) multimodal system, e.g. Visual ChatGPT[27], X-Decoder[32] and InternGPT[18], in which multiple language and vision models are coordinated by a LLM manager/controller to deal with different tasks, 2) end-to-end differentiable multimodal models, e.g. Flamingo[1], BLIP-2[15], MiniGPT-4[30], LLaVA[17], InstructBLIP[6] which input both vision and language tokens into LLM. In this paper, we focus on the end-to-end differentiable multimodal models, which are lightweight and concise for research.

The end-to-end multimodal models contain pre-trained vision and language models and a learnable module to fuse both. Flamingo[1] learns gated cross-attention layers to condition the frozen LLM on visual tokens, demonstrating excellent in-context few-shot learning performance. Li et al. [15] design Q-Former to bridge the image encoder and LLM in a two-stage training strategy, which shows emerging capability of zero-shot instructed image-to-text generation. By leveraging advanced LLMs, i.e. LLaMA[26] and Vicuna[5], multimodal models LLaVA[17] and MiniGPT-4[30] are built by transforming visual tokens to language tokens with only one linear layer, while InstructBLIP[6] learns a Q-Former to bridge vison and language models.

2.2 Multimodal Instruction Tuning

The success of multimodal models, e.g. LLaVA, MiniGPT-4 and InstructBLIP, relies on the high-quality image-text data for finetuning models, which is named visual instruction tuning in [17]. Previous work[9] finds that simply constructing training set based on existing VQA datasets [2, 12] with short answers will degrade the model performance. To boost the performance, Zhu et al. [30] collect 3,500 high-quality image-text pairs by refining their model's outputs using ChatGPT. More natural and reliable responses are produced by finetuning the model on the refined data. Liu et al. [17] for the first time systematically construct a large visual instruction tuning dataset – LLaVA-Instruct-150K. They prompt GPT-4 to generate questions and answers by feeding it image-level captions and object bounding boxes of each image from COCO dataset[16]. Dai et al. [6] collect 26 public datasets including LLaVA-Instruct-150K to construct visual instruction tuning data. However, most of these public datasets contain short questions and answers that focus on visual perception. Beyond above works, we prompt the powerful GPT-4 with rich annotations of image-level captions, region-level descriptions and object bounding boxes that are from Visual Genome[13] and MS-COCO[16]. The generated 3.2M visual instruction data cover diverse tasks of visual perception, reasoning and planing.

There are also some works that contribute multimodal instruction data of videos[14], RGB-D images[14], speech[28], audio[29], et al. For instance, EgoCOT[20] prompts ChatGPT with video captions to generate instructions and responses of detailed embodied planning. MIMIC-IT[14]

collects visual data from multiple datasets, and prompts ChatGPT to generate instruction-response pairs. Most of its data are constructed based on the egocentric videos from E4D dataset[10].

3 Dataset Construction

3.1 Source Data

We build SVIT based on Visual Genome[13] dataset that comprises 108,077 images with dense annotations within each image, including region descriptions, objects, attributes, relationships etc. Since Visual Genome is partially sourced from MS-COCO[16], we also collect captions for images from MS-COCO. Generally, each image in MS-COCO has 5 captions, focusing on the high-level appearance. As an image usually contains rich objects and regions that cannot be completely described in a general caption, Visual Genome serves as a valuable source, offering abundant annotations of the visual details. On average, Visual Genome provides 42 human-generated region descriptions and 21 objects per image, with each region and object located by a bounding box. Leveraging these annotations, we are able to gather thorough and detailed descriptions for all images, which are made up of three key components: (1) the 257,633 captions from MS-COCO; (2) the 3,802,374 object names and their corresponding bounding boxes from Visual Genome; (3) the 5,406,592 region descriptions from Visual Genome.

3.2 Instruction Data Generation

Inspired by LLaVA[17], we design three tasks and prompt the language-only GPT-4 ChatBot to generate the questions and answers accordingly. The prompts are summarized in Fig. F8 in the Appendix. Since GPT-4 demonstrates excellent performance even with zero-shot learning, we do not provide any examples for GPT-4 in order to encourage the innovation and diversity of the generated content.

- **Conversation.** We prompt GPT-4 to design 3 conversations between a person and GPT-4 talking about the image. Each conversation should include 5 question and answer pairs (QAs). The content of the conversation should be logically connected. GPT-4 thinks about the topic first and then generates the conversation according to the topic. The topics can be about the visual perception, reasoning, event planning, etc.
- **Complex reasoning.** 15 complex reasoning QAs about each image are generated using GPT-4. The questions can be asking why things happen that way, suggestions to the people in the image, etc. When providing the answer to a complex question, we prompt GPT-4 to think step by step and include reasoning details in the answer.
- **Detail description.** We use GPT-4 to describe the image in detail. The description may include the people or object appearance, what people or objects are doing, object count, object position, background details, etc.

Fig. 1 illustrates an example input and the GPT-4 output for each task. For rich diversity, we further randomly sample an instruction for detail description task, e.g., "can you describe the image in detail". The complete list of the alternative instructions can be found in Fig. F9 in the Appendix.

3.3 Postprocessing

While most of the GPT-4 generated question-answer pairs are of high quality, some answers occasionally contain unneeded contents. For example, some objects mentioned in the answers are followed by the bounding boxes. Some answers may tell that the information is based on the given "captions" and "descriptions". To remove the unneeded content, we find them based on key words and use GPT-4 to regenerate the responses. In addition, the number of generated conversations or complex reasoning QA pairs may be fewer than the requirement. We also remove them and generate new responses. We use the same procedure to filter the regenerated content until it is satisfying.

3.4 Statistics and Analysis

Statistics. Employing the two-pass procedure, we obtain an extensive collection of data, including 1,565,797 conversation QAs, 1,556,902 complex reasoning QAs, and 106,295 detailed image

Input:

(1) captions:

Street workers in orange jackets outside working with a traffic signal.

The workers are trying to install a broken signal.

people constructing the road and the traffic lights

…

(2) objects:

cone: [0.726, 0.547, 0.758, 0.643]

man: [0.588, 0.387, 0.692, 0.832]

traffic light: [0.16, 0.341, 0.34, 0.571]

…

(3) regions:

orange cone with two white stripes

white vehicle stopped

four men installing a traffic light

Dirt on the street

A traffic light leaning on the truck

All workers has orange vest or coats on

The red and green light is lite up on the signal lights

…

Output:

Conversation:

Topic: Traffic Light Installation

User: How many workers are present in the image?

GPT: There are four workers actively involved in the installation of the traffic light in the image.

User: Can you describe the clothing of the workers?

GPT: The workers are wearing orange jackets, or vests with shirts underneath. The jackets and vests are safety attire, which is typical for street workers to make them more visible to motorists.

User: Are there any safety measures visible in the image?

GPT: Yes, as a safety measure, there are orange cones with white stripes placed on the street near the working area to alert motorists and pedestrians of the ongoing construction.

…

Complex reasoning:

Question: What precautions are the workers taking to ensure their safety?

Answer: The workers are wearing orange jackets, which increase their visibility. Additionally, they've placed orange cones around the area where they are working.

Question: What could be a recommendation to improve safety at this work site?

Answer: Adding more clear signage to alert oncoming traffic of the construction work ahead could be one recommendation to improve safety.

…

Detail description:

In the image, I see a scene where street workers in orange jackets are outside, working with a traffic signal. The workers seem to be trying to install a broken signal. There are several objects and people in the image. I can see cones with two white stripes, indicating the presence of roadwork. There is a man wearing an orange shirt, and he appears to be the main focus of the image. He is accompanied by several other men who are also wearing orange shirts. They are engaged in various activities, such as fixing a street light in the middle of the street. The workers are standing on a street with dirt and orange cones scattered around. In the background, there is a white car and some windows can be seen on it. A green signal is displayed on the traffic light, indicating that it is in working condition. Additionally, there is a truck nearby, and a traffic light seems to be leaning on it. The workers are wearing orange vests or coats, and one of them is wearing a glove on his right hand. Overall, this image captures a group of workers actively engaged in road construction and traffic signal maintenance.

Figure 1: The example input to GPT-4 and the responses for three tasks. Note that the image is only shown here for reference and not provided to GPT-4.
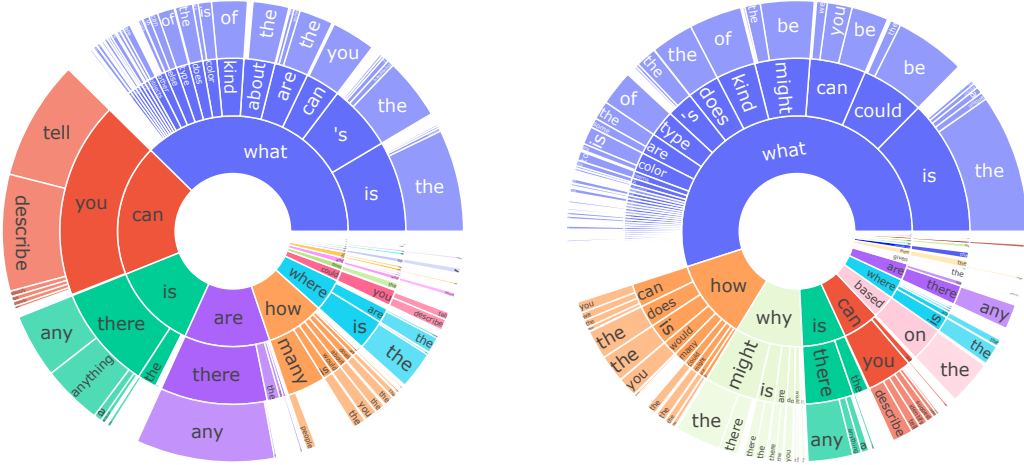
Figure 2: The distribution of question types in *conversations* (left) and *complex reasoning* (right) by the first three words. The angle of each sector represents the proportion of each category.

descriptions. The averaging question and answer lengths are 9.6 and 27.9 words in *conversation* subset, and 12.6 and 26.6 words in *complex reasoning* subset respectively. In contrast, the mean length is 5.7 words per question and 1.8 words per answer in the original Visual Genome. The detailed descriptions in our dataset have 361.5 words on average, while the length of MS-COCO image captions is 11.3. Therefore, the corpus provided by our SVIT is of higher quality.

Distribution. We analyze the distribution of question types in *conversation* and *complex reasoning* tasks by visualizing the distribution of first three words in Fig 2. We can see that "what" questions are the largest category, both in *conversation* (38%) and *complex reasoning* (55%). In the case of *conversation*, question types are diverse, including simple yes-no questions, questions on object details, conditions and functions, etc. Regarding *complex reasoning*, since we explicitly prompt GPT-4 to generate questions that need complex reasoning to answer, we collect a larger proportion of complex questions that commence with "why" (9%) and "how" (11%). Furthermore, most questions starting with "how" are simple object counting questions, i.e. "how many", in existing visual question answering datasets such as Visual Genome[13] and VQA[2], while in SVIT, only 11% of questions starting with "how" are the "how many" questions.

Correctness. To assess the correctness of the generated content, we conduct a manual examination on randomly selected 20 images and corresponding data. In general, around 5% of the questions in the dataset can be provided with a more accurate or satisfying answer. The identified problems can be categorized into three types.

- Errors in original annotations. We construct the visual instruction data based on the manual annotations from Visual Genome and MS-COCO, which may contain errors in their original annotations. For example, in the image depicted in Fig 3a, one caption from MS-COCO incorrectly states, "Three men and one older woman stand near a man who is looking in the mirror with the collar of his white shirt up." Actually, there are only two men and one woman standing near the man looking at the mirror. Similarly, in Fig 3b, the object is labeled as a "little bunny" in the region description, but wrongly referred to a "teddy bear" in the object name in Visual Genome's annotation.

- Correct but not precisely answer the question. As illustrated in Fig 3c, when being asked, "What can be inferred about the maintenance of the boat from the condition of the paint?", the answer states, "The condition of the boat's paint could reflect the level of maintenance, if it's faded or peeling, it may suggest the boat hasn't been maintained well, whereas bright and fresh paint may indicate regular upkeep." Although the answer is correct, it fails to address the question precisely.

- Incorrect answers. In Fig 3d, the generated image description mentions, "Nearby, there's a round center table cluttered with assorted magazines and books, creating a lived-in feel. The table also hosts a yellow rotary telephone, a vintage relic of bygone days." In reality, there

5

(a) Wrong caption in MS-COCO: "Three men and one older woman stand near a man who is looking in the mirror with the collar of his white shirt up."

(b) Wrong object name in Visual Genome: "teddy bear".

(c) The answer discusses how the condition of the boat's paint would reflect the maintenance instead of answering it directly.

(d) The generated answer misunderstands the position of the telephone.

Figure 3: Problematic examples in generated answers.

are two tables in the image and the telephone is placed on a different table in the bottom left corner, though it needs careful observation.

4 Evaluation

To validate the superiority of SVIT dataset, we re-implement the instruction tuning of MiniGPT-4[30] and LLaVA[17] models on SVIT dataset, and then compare to those models tuned on their own instruction datasets. For fair comparison, we follow their training/testing protocols and use the same hyper-parameters. Please refer to the two papers for the implementation details. We believe that better performance can be achieved if improve the training protocol and carefully tune the hyper-parameters. 10% of the images are randomly sampled from SVIT as the held-out testing set for evaluation purpose.

Fine-tuning MiniGPT-4. Following the latest version of MiniGPT-4, the v0 version of Vicuna-13B[5] is used as the language model. Since Zhu et al.[30] only fine-tune MiniGPT-4 on images captions, we follow their setting and only feed the detail description subset of SVIT into the model.

Fig. 4 illustrates the comparison between models trained on their 3,500 instruction data and that trained on our SVIT. The image portrays a woman holding a birthday cake, accompanied by a child. When asked about the atmosphere of the scene, the model trained on MiniGPT-4's dataset fails to analyze the image correctly. It imagines the expression of people out of the image and has the hallucination that the room is decorated with balloons and streamers, which have little evidence. In contrast, the SVIT-trained model comprehends the image correctly. It gathers information from different perspectives such as the birthday cake, people's spirits and clothes to draw its conclusion. When asked about the appearance of the cake, the model trained on SVIT generates a detailed and precise answer – "a yellow-frosted round cake", while the model trained on MiniGPT-4's dataset is not able to make a description, instead, it imagines the appearance of the cake based on the scene.

Fine-tuning LLaVA. LLaVA supports both Vicuna-7B and Vicuna-13B as the language model. In our experiment, we use Vicuna-7B to accelerate the training process.

Training LLaVA model on SVIT dataset enhances the model's capabilities. As demonstrated in Fig 5a, the input image depicts a scene of food arranged on a table. The model trained on LLaVA

What is the atmosphere of the scene?

MiniGPT-4: The atmosphere of the scene is joyful and celebratory, as the mother is carrying a birthday cake with candles for her child's birthday party. The mother and child are surrounded by a group of people, who are looking on with happy expressions. There are smiles and laughter in the air, and the mood is jovial.

The room is decorated with balloons and streamers, adding to the festive atmosphere. The sunlight coming through the window adds a natural touch to the scene, making it feel welcoming and cheerful.

SVIT: Based on the provided image, the atmosphere of the scene is celebratory. A woman is shown holding a birthday cake, indicating that there is a birthday celebration happening. People are present, including a child, and they all appear to be in good spirits. The woman is wearing casual clothing, which suggests that the event is taking place in a relaxed, informal setting.

Overall, the atmosphere of the scene appears to be lively and joyous, with everyone gathered together to celebrate a special occasion.

What does the cake look like?

MiniGPT-4: I'm sorry, as I am a text-based AI language model, I am not able to see or describe images. However, based on the description of the scene, the cake may be a birthday cake with candles on top, possibly decorated with icing and sprinkles. The size and shape of the cake are not specified, so it is unclear what it looks like.

SVIT: Based on the provided image, the birthday cake looks like a yellow-frosted round cake. The frosting is applied evenly over the cake, creating a smooth and glossy surface. The frosting is a pale yellow color, which contrasts with the white of the cake and provides an eye-catching look.

It is difficult to make out the specific design or decorations on the cake, but it appears to have a simple and classic appearance. Given the context of the image, it is likely that the cake is meant to serve as a focal point for the birthday celebration, and it fulfills this role with its vibrant color and frosting.

The overall presentation of the cake is attractive and inviting, which is fitting for a birthday celebration where everyone is gathering to enjoy sweets and good company.

Figure 4: Question answering. "MiniGPT-4" and "SVIT" denote the two models trained on datasets of MiniGPT-4 and our SVIT respectively. The model trained on SVIT is able to answer the question more precisely and reduce the hallucination.

(a)

(b)

Figure 5: Image description. "LLaVA" and "SVIT" denote the two models trained on datasets of LLaVA and our SVIT respectively. The model trained on SVIT can describe abundant details accurately.
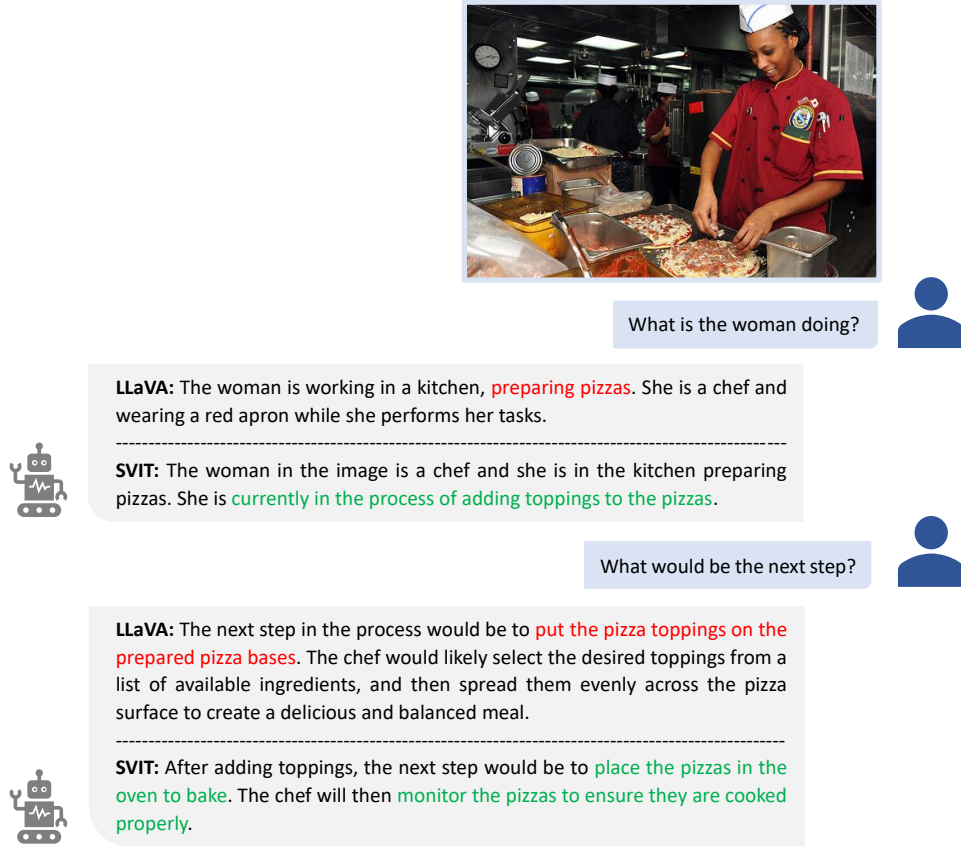
Figure 6: Event planning. "LLaVA" and "SVIT" denote the two models trained on datasets of LLaVA and our SVIT respectively. The model trained on SVIT captures the exact action the woman is doing and proposes the subsequent steps logically.

dataset inaccurately states that there are mushrooms on the steak and a spoon near the edge of the table. In contrast, the SVIT-trained model successfully figures out the content in the image, not only recognizing the main dish, but also identifying the checkered tablecloth and the silver bowl of sauce in the background. Similarly, in Fig 5b, the model trained on LLaVA dataset inaccurately mentions the small figures of people visible in the background. Whereas, the model trained on SVIT analyzes the scene by considering details, such as the man's clothing, accessories and objects present in the background. Furthermore, the model can precisely figure out the small toys.

To evaluate the ability of planning, we ask both models what is happening in the image and prompt the models to plan the subsequent steps. In Fig 6, the input image shows that a woman is adding toppings to the pizzas. The model trained on SVIT accurately captures the exact action, while the model trained on LLaVA dataset only roughly identifies that the woman is preparing a pizza. In addition to the precise understanding of the event, the model trained on SVIT logically plans the next steps, i.e. baking the pizzas in the oven and monitoring them.

Regarding conversational performance, Fig 7 demonstrates a conversation between the user and agent discussing an image. The SVIT-trained model correctly identifies the number of cars covered in snow, whereas the model trained on LLaVA dataset fails to do it. When asked about the suggestions for the driver, the SVIT-trained model offers a comprehensive assessment, taking multiple factors into consideration, such as the icy road, tire treads, blurry windows and mirrors.
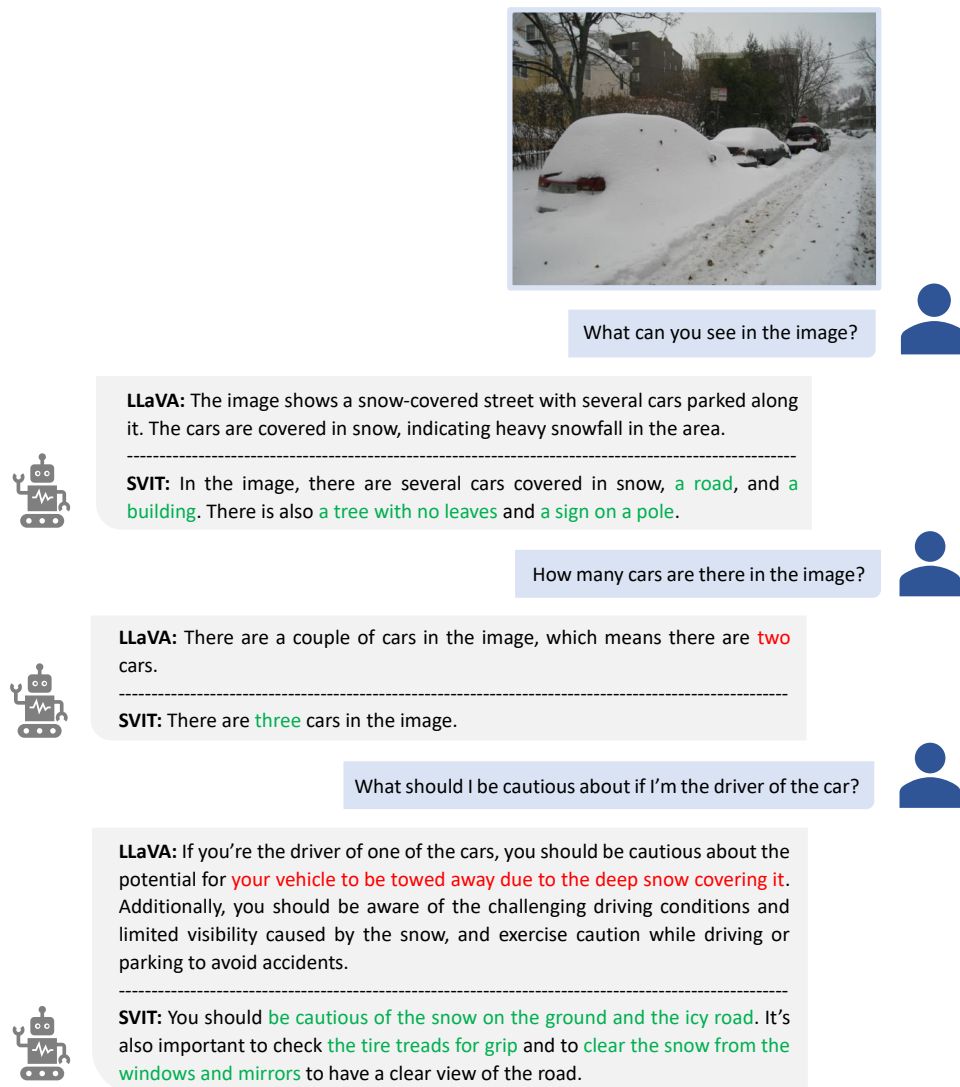
What can you see in the image?

LLaVA: The image shows a snow-covered street with several cars parked along it. The cars are covered in snow, indicating heavy snowfall in the area.

SVIT: In the image, there are several cars covered in snow, a road, and a building. There is also a tree with no leaves and a sign on a pole.

How many cars are there in the image?

LLaVA: There are a couple of cars in the image, which means there are two cars.

SVIT: There are three cars in the image.

What should I be cautious about if I'm the driver of the car?

LLaVA: If you're the driver of one of the cars, you should be cautious about the potential for your vehicle to be towed away due to the deep snow covering it. Additionally, you should be aware of the challenging driving conditions and limited visibility caused by the snow, and exercise caution while driving or parking to avoid accidents.

SVIT: You should be cautious of the snow on the ground and the icy road. It's also important to check the tire treads for grip and to clear the snow from the windows and mirrors to have a clear view of the road.

Figure 7: Conversation. "LLaVA" and "SVIT" denote the two models trained on datasets of LLaVA and our SVIT respectively. The model trained on SVIT identifies the number of cars covered in snow correctly and proposes suggestions considerately.

10

5 Conclusion

In this paper, we scale up visual instruction tuning by presenting a large-scale dataset – SVIT that contains in total 3.2 million instruction tuning data. The instructions and responses are generated by prompting GPT-4 with abundant manual annotations of images. Experiments show that training multimodal models on our large dataset achieves better performance in tasks of visual perception, reasoning and planing, compared to training models on other small datasets.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[11] 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.

[12] 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.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[14] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[18] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[20] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[21] 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.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[25] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[27] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[28] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.

[29] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[30] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[31] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

[32] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

Appendix

A Prompts

Based on the captions, object bounding boxes and region descriptions of images, we design three tasks and prompt GPT-4 to respond accordingly. We do not include the bounding boxes of region descriptions in the input data, since the context length may exceed the limit of GPT-4 in many cases. The prompts share the same paragraph describing the input data at the beginning and then differ in task description, which are summarized in Fig F8.

B Instructions for Detail Description

Fig. F9 shows the instructions for detail description. We prompt GPT-4 to generate different ways of saying "can you describe the image in detail" and accumulate all the instructions. For each image, we randomly sample one from the list as instruction.

You are an AI visual assistant that can analyze a single image. The information of the image is made up of three parts:
(1) "captions": If it is not empty, it contains five sentences, describing the image you are observing.
(2) "objects": It contains multiple lines, each describing an object of the same image you are observing. Every line is made up of an object name and its bounding box. The bounding box is in the form of [x1, y1, x2, y2]. The values are float numbers normalized from 0 to 1, corresponding to the top left x, top left y, bottom right x, and bottom right y.
(3) "regions". It contains multiple lines, each describing a region of the same image you are observing.

Conversation:
Design three conversations between you and a person asking about this image. A conversation should include five question and answer pairs. The content within the conversation should be logically connected. You can think about the topic first and then generate the conversation according to the topic. The topic can be the visual content of the image (including the object types, counting the objects, object actions, object locations, relative positions between objects, etc.), the background knowledge of the objects, the events happening in the image, event planning, etc. In the conversation, you are called "GPT". The person talking with you is called "User".

Ask diverse questions and give corresponding answers. Only include questions that have definite answers. The answer should be in a tone that a visual AI assistant is seeing the image and answering the question. The length of the answer would better be within 50 tokens.

When using the information from the description, do not mention that the information source is the description. When using the information from the object bounding box, do not mention that the information comes from the bounding box as well. Always answer as if you are directly looking at the image.

Complex reasoning:
Create 15 plausible question and answer pairs about the image with provided information.

The question requires commonsense knowledge about the scene and can only be answered with image provided. Avoid asking questions that can be answered with commonsense knowledge alone. Avoid proposing questions that can be answered with simple visual understanding like asking about object type and color. Do not give too many details about the visual content of the image, so one has to figure it out first to answer the question correctly. The question can be asking why things happen that way, suggestions to the people in the image, etc. When providing the answer for complex questions, think step by step and include reasoning details.

When using the information from the description, do not mention that the information source is the description. When using the information from the object bounding box, do not mention that the information comes from the bounding box as well. Always answer as if you are directly looking at the image.

Desired format:
Question: ...
Answer: ...
Question: ...
Answer: ...

Detail description:
The task is describing the image in detail. Though you do not receive the pixel data of the image, utilize above textual information to think about the image and describe as if you are directly looking at the image. The description can include what people or objects are doing, object appearance, object count, object position, background details, etc. Only describe the things that you are sure about.

When using the information from the description, do not mention that the information source is the description. When using the information from the object bounding box, do not mention that the information comes from the bounding box as well.

Figure F8: The prompts of conversation, complex reasoning and detail description to GPT-4.

Instructions for detail description

- Can you provide a comprehensive description of the image?
- Elaborate on the details of the image.
- What are the specifics visible in the image?
- Could you offer an in-depth analysis of the image?
- Can you depict the image with precise detail?
- Give a detailed account of the image.
- Explain the image in meticulous detail.
- Can you portray the image in words?
- Give a thorough narrative of the image.
- Please provide an intricate breakdown of the image.
- Offer a complete interpretation of the image.
- Delve into the particulars of the image.
- Explain all the nuances you observe in the image.
- Provide a detailed commentary on the image.
- Illustrate the image in depth using your words.
- Could you give a blow-by-blow description of the image?
- Go into detail about the different elements of the image.
- Can you dissect the image and describe each element in detail?
- Detail the contents of the image extensively.
- Can you provide an in-depth explanation of the image?
- Provide a comprehensive overview of the image.
- Break down the elements of the image in detail.
- Can you expound upon the features of the image?
- Offer an exhaustive description of the image.
- How would you illustrate the image in words?
- Please convey the image's details verbally.
- Can you detail the contents of the image?
- Narrate what you see in the image in depth.
- Kindly provide a meticulous commentary on the image.
- Share an extensive description of the image.
- Could you interpret the image in a detailed manner?
- Present a detailed report of the image's features.
- Can you provide an intricate depiction of the image?
- Disclose every detail you see in the image.

Figure F9: Instructions for detail description.