Supplementary Material of "SQ-LLaVA: Self-Questioning for Large Vision-Language Assistant"

Guohao Sun[1], Can Qin[2], Jiamian Wang[1], Zeyuan Chen[2], Ran Xu[2], and Zhiqiang Tao[1]

[1] Rochester Institute of Technology, Rochester, NY, US
[2] Salesforce AI Research, CA, US

1 Overview

We provide more results, discussions, and qualitative examples of the proposed SQ-LLaVA in **13B** scale.

1. Ablation study of SQ-LLaVA trained on two visual instruction dataset (Section 2.1).
2. Evaluation on an additional vision-language benchmark (Section 2.2).
3. Qualitative examples of image description and self-questioning (Section 3).

2 Zero-shot Multilingual Capability

2.1 Ablation Study

In Table 1, we conduct experiments with different architecture designs and training strategies on five benchmarks for SQ-LLaVA-13B. For a fair comparison, we train the baseline models on our local machine with the same training recipe of LLaVA-LoRA [4]. Specifically, we present the baseline model, our full model, and three ablated models by removing one component each time. We adopt the dataset [4] with 558k for pre-training (PT) and 665k for fine-tuning (FT). As compared, self-questioning (SQ) achieves 0.6% improvement in average accuracy, indicating the effectiveness of visual self-questioning in improving visual language understanding. Besides, we introduce the prototype extractor (Proto) to enhance visual representation, achieving 0.6% improvement in average accuracy. With all three components incorporated, we observe a 2.5% improvement in average accuracy among five benchmarks.

As shown by the bottom block of Table 1, we conduct experiments with the same ablation settings but with a larger scale of the visual instruction data [1] (*i.e.*, for both PT and IT). Overall, SQ-LLaVA achieves 2.5% improvement over the baseline model after training on the smaller dataset and achieves 3.2% improvement after training on the larger dataset. In Table 1, we observe that the proposed SQ-LLaVA achieves significant improvement over the baseline models

Table 1: Ablation study of training strategy on visual instruction tasks. All models are in **13B** scale with three components of ViT-LoRA (V-LoRA), self-questioning (SQ), and prototype extractor (Proto). We provide experiments on two instruction datasets [1,4] with different pre-training (PT) and instruction tuning (IT) data scales.

PT	IT	V-LoRA	SQ	Proto	VizWiz	SQA^I	VQA^T	POPE	$LLaVA^W$	Avg.
558K	665K	✗	✗	✗	53.9	68.2	60.2	86.7	69.5	67.7
		✓	✗	✓	54.5	70.4	60.6	87.7	71.6	69.0
		✓	✓	✗	54.8	69.4	60.7	87.5	72.8	69.0
		✗	✓	✓	54.1	68.6	60.1	88.0	71.0	68.4
		✓	✓	✓	54.6	69.8	60.2	87.7	74.6	69.4
1200K	700K	✗	✗	✗	53.6	70.1	62.0	87.0	76.4	69.8
		✓	✗	✓	56.8	71.0	61.9	87.4	79.6	71.3
		✗	✓	✓	56.2	70.2	61.4	86.9	83.6	71.7
		✓	✓	✓	58.2	71.7	62.0	87.4	80.7	72.0

on the LLaVA (in-the-wild) benchmark by 7.3% and 5.6%. This indicates SQ-LLaVA's capability in more challenging tasks, generalizability to novel domains, and robustness to different prompts.

Table 2 shows the experimental results by applying self-questioning (SQ) on two different backbone models, where SQ consistently improves the performance. The special token [vusr] is the key design that instructs the LLM on when to perform SQ in training/inference, applied to general benchmarks. Its ablation is equivalent to not using SQ.

Table 2: Ablation study of self-questioning (SQ).

Method	LLM	VizWiz	SQA^I	VQA^T	POPE	$LLaVA^W$	Avg.
LLaVA	7B	49.4	68.4	58.2	86.5	67.1	65.9
LLaVA + SQ	7B	**56.8**	**68.5**	**58.6**	**87.7**	**67.8**	**67.9**
Sharegpt4V	7B	51.5	68.9	58.9	86.8	72.1	67.6
Sharegpt4V + SQ	7B	**53.9**	**69.6**	**59.1**	**87.3**	**72.4**	**68.5**

2.2 Evaluation on An Additional Benchmark

We evaluate SQ-LLaVA on one recent benchmark MME [3]. MME is designed to assess the performance of vision-language models across a wide range of tasks that require both perception and cognition abilities. In Table 3, we provide a comparison between SQ-LLaVA with other methods. As can be seen, SQ-LLaVA achieves the highest cognition score among all baseline models, indicating a strong performance when handling the commonsense reasoning, numerical calculation, text translation, and code reasoning tasks.

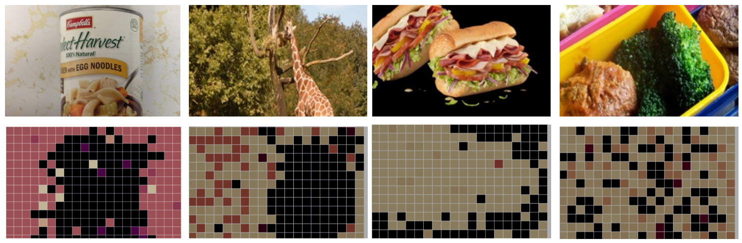


Fig. 1: Prototype clustering on visual tokens (image patches). Similar patterns are grouped in the first three, yet failed in the last.

Table 3: Comparison with SOTA methods on recent evaluation benchmarks. The best results are **bold** and the second-best results are underlined.

Model	MME^P	MME^C
InstructBLIP [2]	1212.8	291.8
LLaVA-v1.5 [4]	1531.3	295.4
ShareGPT4V [1]	**1618.7**	303.2
SQ-LLaVA	1568.2	**326.8**

3 Qualitative Analysis of SQ-LLaVA

3.1 Prototype Visualization

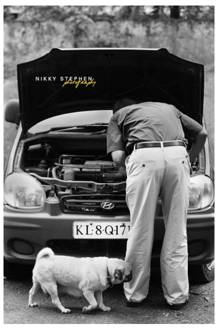
Fig. 1 shows the prototype clustering result over visual tokens. While it groups similar patterns on the first three images, the prototype fails with complex backgrounds due to its unsupervised nature (the lack of prototype labels).

3.2 Reliable Detailed Image Description

The detailed image description examples demonstrate that SQ-LLaVA can capture low-level visual information within the image, such as "KL8-Q17" in Fig. 2 and "number 186 206-7" in Fig. 3. Also, SQ-LLaVA demonstrates a better reliability. In Fig. 3, the image shows a complex environment with various objects, which makes it easy for the models to generate wrong descriptions such as "total five people" (LLaVA-v1.5) and "faces blurred" (ShareGPT4V). In contrast, SQ-LLaVA describes uncertainty information with a prefix hedge word such as "likely", "possibly", and "perhaps", achieving higher reliability in a real-world application.

3.3 Visual Self-questioning

In Fig. 4, Fig. 5, and Fig. 6, we show visual self-questioning and answering of SQ-LLaVA. As can be seen, SQ-LLaVA asks meaningful questions based on comprehensive visual understanding. Specifically, the self-asked questions always involve one or more main objects within the image and a relatively divergent problem, where the main objects rely on visual perception, and the meaningful problem requires vision-language alignment. Unlike question-answering with direct instruction, self-questioning encourages the model to discover all possible combinations between visual and semantic information.



Detailed Image Description (Model Trained on LLaVA-v1.5 Data)

SQ-LLaVA-13B (Ours): The image features a man working on a car, with the hood open, while a small white dog stands nearby, attentively watching the man's actions. The man appears to be focused on fixing the car, possibly a Hyundai, as he leans over the hood. The dog is positioned close to the man, showing interest in the ongoing activity.

LLaVA-v1.5-13B: The image features a man standing next to a car with its hood open, working on the engine. A small white dog is standing close to the man, attentively watching him as he works. The dog appears to be curious about the man's activity. The car is parked in a driveway, and the man is focused on the task at hand. The scene captures the interaction between the man and the dog, with the dog's interest in the man's actions.

Detailed Image Description (Model Trained on ShareGPT4V Data)

SQ-LLaVA-13B (Ours): In this black and white photograph, a man is engrossed in the task of working on a car. He is dressed in a casual outfit consisting of a t-shirt and jeans, and he's wearing a belt. The car, which is the main subject of the image, is parked on a grassy field. The hood of the car is open, revealing the engine, indicating that the man is likely performing some maintenance or repair work. The car's license plate reads \"KL8-Q17\", providing a unique identifier for the vehicle. The man is standing on the right side of the car, leaning over the open hood, suggesting that he is deeply involved in his task. Adding a touch of life to the scene, a small dog is present in the foreground. The dog, standing on the left side of the car, is looking up at the man, possibly curious about his actions. The dog's presence adds a sense of companionship to the scene, suggesting that the man and the dog might share a bond. Overall, the image captures a moment of everyday life, where a man is taking care of his car, and a small dog is observing the scene. The black and white color scheme adds a timeless quality to the image, highlighting the contrast between the man, the car, and the dog.

ShareGPT4V-13B: In this black and white photograph, a man is engrossed in the task of inspecting the engine of a car. The car, a Hyundai, is parked on a grassy field, its hood open to reveal the intricate details of its engine. The man, dressed in a casual shirt and pants, is bent over the engine, his attention completely absorbed by the task at hand. A few feet away from the man, a white dog stands attentively, its gaze fixed on the man and the car. The dog's position suggests it might be the man's pet, accompanying him on this outdoor adventure. The photograph is taken from a low angle, adding a sense of depth and perspective to the scene. The man and the dog are the main subjects of the image, their actions and positions telling a story of a day spent outdoors, perhaps on a road trip or a simple day out in nature. The black and white color scheme adds a timeless quality to the image, evoking a sense of nostalgia.

Fig. 2: A qualitative evaluation of detailed image descriptions from four models. We highlight the words and sentences that represent how each model describes the main object in the image. We use underlined word to indicate how SQ-LLaVA describe uncertain information.

Detailed Image Description (Model Trained on LLaVA-v1.5 Data)

SQ-LLaVA-13B (Ours): The image depicts a red train traveling down the tracks, pulling into a train station. There are several people waiting on the platform, some of them carrying backpacks and handbags. A few individuals are standing closer to the train, while others are positioned further away. In addition to the people and the train, there are two benches visible on the platform, one near the center and the other towards the right side. A suitcase can also be seen on the platform, <u>likely</u> belonging to one of the passengers.

LLaVA-v1.5-13B: The image features a red train traveling down the tracks, pulling into a station. There are several people waiting on the platform, some of them carrying handbags and backpacks. A total of five people can be seen in the scene, with one person standing closer to the left side of the platform, another person in the middle, and three others on the right side. The train occupies a significant portion of the image, stretching from the left to the right side of the platform. The people on the platform are likely waiting for the train to come to a complete stop before boarding.

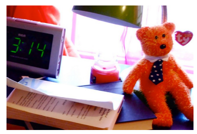
Detailed Image Description (Model Trained on ShareGPT4V Data)

SQ-LLaVA-13B (Ours): The image captures a moment at a train station. A red train, numbered 185 206-8, is stationed on the tracks, ready to embark on its journey. The train's vibrant red color stands out against the more muted tones of the station. The platform, constructed of solid concrete, is populated with people, their forms slightly blurred, suggesting movement. They are waiting, <u>perhaps</u> for the train to depart or for someone to arrive. The station itself is a blend of modern and traditional architecture. The roof, a modern design, is supported by sturdy metal beams. In contrast, the walls are constructed of brick, adding a touch of rustic charm to the scene. The sky above is a clear blue, dotted with fluffy white clouds, suggesting a pleasant day. The image is a snapshot of everyday life, capturing the anticipation and routine of travel.

ShareGPT4V-13B: The image captures a bustling scene at a train station. A red train, bearing the number 186 206-7, is stationed on the tracks, ready to embark on its journey. The train's vibrant red color stands out against the more muted tones of the station. The platform, constructed of solid concrete, is sheltered by a metal roof, providing protection from the elements for waiting passengers. A handful of people can be seen on the platform, their faces blurred, adding a sense of motion to the scene. In the background, the station continues with more tracks and buildings, painting a picture of a busy transportation hub. The sky above is a clear blue, suggesting a bright and sunny day. The image is a snapshot of everyday life, capturing the anticipation and movement inherent in such a setting.

Fig. 3: A qualitative evaluation of detailed image descriptions from four models. We highlight the words and sentences that represent how each model describes the main object in the image. We use <u>underlined</u> word to indicate how SQ-LLaVA describe uncertain information.

User



SQ-LLaVA: What kind of environment does the desk appear to be in?

The desk appears to be in a comfortable and inviting environment, likely a home office or study room. The presence of the teddy bear, the open book, and the candle suggests that this space is used for work, study, or relaxation. The pink curtains in the background add a touch of warmth and softness to the scene, further enhancing the cozy atmosphere. :SQ-LLaVA

SQ-LLaVA: What color is the teddy bear in the image?

The teddy bear in the image is orange. :SQ-LLaVA

SQ-LLaVA: Is it possible for this scene to be both "colorful" and "warm"?

Yes, it is possible for this scene to be both "colorful" and "warm." The presence of the red candle, the pink curtains, and the teddy bear with its blue tie and white polka dots all contribute to a warm and inviting atmosphere. The colors in the scene are vibrant and varied, which adds to the overall visual appeal and creates a sense of warmth and comfort. :SQ-LLaVA

SQ-LLaVA: Can you tell what type of workspace this is? Home office, academic, etc.?

Based on the objects present in the image, it's difficult to definitively determine the type of workspace. However, the presence of a book, a teddy bear, and a candle suggests that it could be a home office or a personal study area. The combination of these objects indicates a comfortable and relaxed environment, which is often associated with home offices or personal study spaces. :SQ-LLaVA

Fig. 4: Visual self-questioning and answering of SQ-LLaVA.

User

SQ-LLaVA: What are some of the main elements or object types in the image?

The main elements or object types in the image include a skateboarder, a red crate, a bicycle, a building, and a sign. :SQ-LLaVA

SQ-LLaVA: What is the occupation of the person in the image?

The person in the image is a skateboarder, performing a trick on a red crate. :SQ-LLaVA

SQ-LLaVA: What is the activity taking place in the photo?

The activity taking place in the photo is skateboarding, with a skateboarder performing a trick on a red crate. :SQ-LLaVA

SQ-LLaVA: What type of environment does this photo seem to depict? Urban or rural?

The photo seems to depict an urban environment, as evidenced by the presence of a city sidewalk, a building, and a bicycle. The skateboarder performing a trick on a crate also suggests that this is a city setting where skateboarding is a popular activity. The image captures the energy and dynamism of city life, with the skateboarder's action being a prime example of the excitement and creativity that urban environments often foster. :SQ-LLaVA

Fig. 5: Visual self-questioning and answering of SQ-LLaVA.

User

SQ-LLaVA: Can you please elaborate more on the three types of donuts present on the plate?

SQ-LLaVA: Certainly! The image features three distinct types of donuts, each with its own unique characteristics and flavors.
1. Chocolate Donut: This donut is covered in a rich, dark chocolate glaze, giving it a decadent and indulgent appearance. The chocolate glaze is likely to have a sweet and slightly bitter taste, making it a popular choice for those who love chocolate.
2. Glazed Donut: This donut is coated with a light, shiny glaze that gives it a glossy appearance. The glaze is typically made from a mixture of powdered sugar, water, and flavorings, and it adds a sweet and slightly sticky texture to the donut. Glazed donuts are a classic favorite and are often enjoyed for their simple yet satisfying taste.
3. Twisted Donut: This donut has a unique twisted shape, which sets it apart from the other two donuts. The twisted design not only makes the donut visually interesting but also adds a different texture to the bite. The twisted donut could be filled with various sweet fillings, such as jam, cream, or custard, or it could be a plain, unfrosted donut.
In summary, the three donuts on the plate represent a diverse range of flavors and textures, from the rich chocolate glaze to the light and shiny glaze, and the unique twisted shape. This variety makes the plate an appealing and enjoyable treat for anyone with a sweet tooth.

SQ-LLaVA: Is there any tea or beverage on the plate?

SQ-LLaVA: No, there is no tea or beverage on the plate. The plate only contains two chocolate donuts and a cinnamon roll.

Fig. 6: Visual self-questioning and answering of SQ-LLaVA.

References

1. Chen, L., Li, J., wen Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: ArXiv (2023)
2. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B.A., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
3. 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. In: ArXiv (2023)
4. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: ArXiv (2023)