







SQ-LLaVA: Self-Questioning for Large Vision-Language Assistant

Guohao Sun¹, Can Qin², Jiamian Wang¹, Zeyuan Chen²,
Ran Xu², and Zhiqiang Tao¹

¹ Rochester Institute of Technology, Rochester, NY, US

² Salesforce AI Research, CA, US

Abstract. Recent advances in vision-language models have shown notable generalization in broad tasks through visual instruction tuning. However, bridging the gap between the pre-trained vision encoder and the large language models (LLMs) becomes the whole network’s bottleneck. To improve cross-modality alignment, existing works usually consider more visual instruction data covering a broader range of vision tasks to fine-tune the model for question-answering, which, however, is costly to obtain and has not thoroughly explored the rich contextual information contained in images. This paper first attempts to harness the overlooked context within visual instruction data, training the model to self-supervised “learning” how to ask high-quality questions. In this way, we introduce a novel framework named SQ-LLaVA: Self-Questioning for Large Vision-Language Assistant. SQ-LLaVA exhibits proficiency in generating flexible and meaningful image-related questions while analyzing the visual clue and prior language knowledge, signifying an advanced level of generalized visual understanding. Moreover, fine-tuning SQ-LLaVA on higher-quality instruction data shows a performance improvement compared with traditional visual-instruction tuning methods. This improvement highlights the efficacy of self-questioning techniques in achieving a deeper and more nuanced comprehension of visual content across various contexts. Our code is available at <https://github.com/heliossun/SQ-LLaVA>.

Keywords: Vision-Language Understanding · Multi-modal LLM · Instruction Tuning

1 Introduction

The recently emerging large vision-language methods, such as large language-and-vision assistant (LLaVA) and its variants [3, 5, 21, 22, 54], fine-tune large language models (LLM) [3, 54, 57] on visual instruction data [18, 55, 56, 58] to realize diverse open-world multimodal understanding, demonstrating a surprising efficacy of *visual instruction tuning* – the LLM learns to perform complex vision tasks by conditioning on a prompt containing image and text clues. Existing visual instruction data is mainly built upon conversations (e.g., ChatGPT/GPT4-V [1]), consisting of images and multiple question-answer pairs. Building high-quality visual instruction data usually requires images and texts from different tasks to

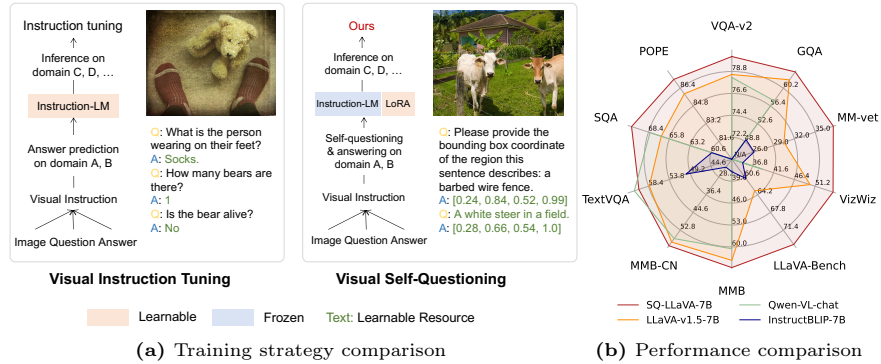


Fig. 1: (a) Comparison between visual instruction tuning and visual self-questioning (ours) for vision-language assistant. (b) The proposed SQ-LLaVA achieves state-of-the-art performance on 9 out of 10 tasks compared with other open-ended models.

generate diverse questions, such as “Please provide the bounding box coordinate of the region this sentence describes A dead leaf on the ground” for the object detection task. Empirically, by increasing the diversity of questions, LLaVA has achieved better performance on GQA and VizWiz tasks (26% and 45% over previous state-of-the-art methods [8]). This evidence strongly suggests the advantage of training models on a broad spectrum and diverse array of tasks for enriching general vision-language understanding.

LLaVA [21] model family usually consists of a pre-trained vision encoder (e.g., CLIP-ViT [33]), a large generative language model (LLMs like Vicuna [57], LLaMA [54], Qwen [3], etc.), and a vision-to-language projector implemented by a few linear layers. However, the modality gap between the pre-trained vision encoder and the language model restricts both sides’ generalization ability and feature representation. To overcome this challenge, various techniques have been proposed to align the vision and language domains, which could be roughly categorized into three groups: 1) build a more robust image feature extractor [8, 54, 58], 2) collect more high-quality training data [5, 8, 21, 58], and 3) fully fine-tune the vision and language models simultaneously during the pre-training stage [5]. While these methods have shown good progress in mitigating the domain gap, they inevitably bring higher computational costs and more expensive data collection and may also require sophisticated manual designs as well as heavy annotating efforts. Moreover, images generally encompass rich information, including color, context, and the relationships between objects, but most existing visual instruction datasets capture only a fraction. We posit that leveraging such under-explored knowledge could significantly aid in vision-language understanding. In this study, we propose a new *visual self-questioning* approach (Fig. 1a) by training the LLM to ask questions and discover vision clues without collecting extra data from other sources.

Unlike existing visual instruction tuning methods that focus solely on answer prediction, the proposed visual self-questioning aims to extract relevant question

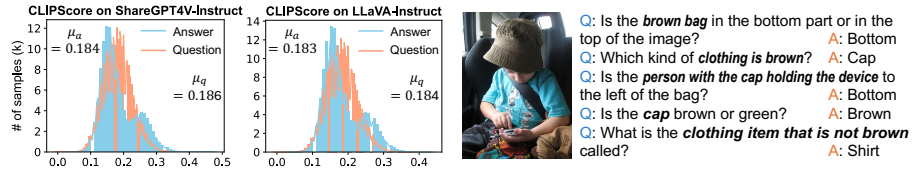


Fig. 2: *Left:* Questions spread more samples around a higher mean CLIPScore than answers. *Right:* Example of highly image-relevant questions within the visual instructional dataset for training a Vision-Language assistant.

context. As illustrated in Fig. 2 *right*, the questions could contain more image-related information than the answers, inspiring us that aligning images and related questions may further improve the model’s vision-language understanding capacity. Quantitatively, we compute the CLIPScore [33] (a higher value means better visual-text relevance) for all the image-question/-answer pairs on two visual instruction datasets (LLaVA-instruct [21] and ShareGPT4V-instruct [5]). By comparison, the questions’ mean CLIPScore $\mu_q=0.184$ is larger than the answers’ $\mu_a=0.183$ on LLaVA-instruct; $\mu_q=0.186$ is larger than the answers’ $\mu_a=0.184$ on ShareGPT4V-instruct. Fig. 2 *left* shows the CLIPScore distribution, indicating questions have a similar and even better visual relevance than the answers. Therefore, questions in current visual instruction data can be used to fine-tune instructional LLMs owing to their diverse semantic information.

This work introduces a self-questioning LLaVA, namely SQ-LLaVA, to fully utilize questions within the instruction data as an additional learning resource for training instructional LLMs and empowering the model’s curiosity (questioning ability). To efficiently align the vision and language domains, we apply LoRAs [12] to optimize both the vision encoder and the instructional LLM within the SQ-LLaVA. Plus, we develop a prototype extractor to enhance visual representation by leveraging learned clusters with meaningful semantic information to improve vision-language alignment further. Extensive experiments demonstrate that SQ-LLaVA surpasses existing visual instruction tuning methods in general vision understanding (see Fig. 1b). We summarize the contributions as follows.

- We propose a novel training technique, visual self-questioning for vision-language assistants (SQ-LLaVA), by leveraging highly relevant question contexts in instruction data. This SQ learning task promotes instructional LLMs to understand the relationship between images and questions, enhancing vision-language alignment without needing new data collection.
- We design and develop a lightweight tuning architecture for SQ-LLaVA, consisting of ViT-LoRA, LLM-LoRA, and a prototype extractor. The prototype extractor enhances vision embeddings, while ViT-LoRA and LLM-LoRA efficiently align vision and language domains during training.
- Extensive experimental results show that the proposed SQ-LLaVA leads to better performance in several tasks, including visual question-answering, visual instruction benchmarks, and zero-shot image captioning.

2 Related Work

2.1 Instruction Tuning

Instruction tuning emerged as a pivotal methodology within the realm of natural language processing (NLP), facilitating Large Language Models (LLMs) such as GPT-3 [4], PaLM [7], and LLaMA [42] to interpret and execute human language instructions across a spectrum of NLP tasks. This approach diverges from traditional fine-tuning mechanisms by incorporating a specialized data structure, termed instruction-following data [50], which is instrumental in the fine-tuning process of LLMs. Generally, there are two main categories regarding instruction tuning methods – 1) closed-domain and 2) open-domain. The closed-domain instruction tuning [35, 50, 51] studies engaged LLMs with a comprehensive assortment of publicly accessible datasets and subsequently assessed their performance across diverse NLP tasks [44]. The empirical evidence from these inquiries consistently indicated that integrating varied NLP task instructions significantly augments the LLMs’ efficacy in navigating novel tasks. Nonetheless, LLMs calibrated with such closed-form instructions exhibited limitations in real-world user scenarios, prompting the development of an alternative approach. To address these constraints, the concept of open-domain instruction tuning [30, 49] is conceived. OpenAI pioneered this approach by employing human annotators to compile a corpus of real-world question-answer datasets. These datasets form the foundation for training a reward model through reinforcement learning methodologies. The trained reward model then functions as a supervisory mechanism for further training instruction-oriented language models, such as InstructGPT [31] and Vicuna [57]. This innovation marks a significant advancement in the field, aiming to bridge the gap between LLM performance and real-world applicability by leveraging instruction data derived from authentic user interactions.

2.2 Large Vision Language Models

As the field of LLMs and instruction tuning undergoes rapid advancements, the academic research community is progressively focusing on integrating visual information into LLMs to facilitate visual instruction tuning. This emerging area of research has witnessed the development of various methodologies [3, 5, 21, 22, 54] based on foundational vision-language models [19, 33, 40, 47] and diverse LLM architectures [3, 43, 57]. In particular, LLaVA [22] pioneers the integration of an LLM with a CLIP vision encoder to construct a vision language model, demonstrating remarkable capabilities in image-text dialogue tasks through pretraining alignment strategies and targeted instruction tuning. Subsequent investigations have sought to refine visual instruction tuning by enhancing the quality and variety of the datasets used during the pre-training and fine-tuning phases. Building upon these advancements, recent studies like LLaVA-v1.5 [21] and ShareGPT4V [5] have achieved notable success in general vision-language comprehension, showcasing their ability to undertake complex question-answering tasks. This progression underscores the importance of sophisticated data handling and model tuning strategies in developing effective vision-language models.

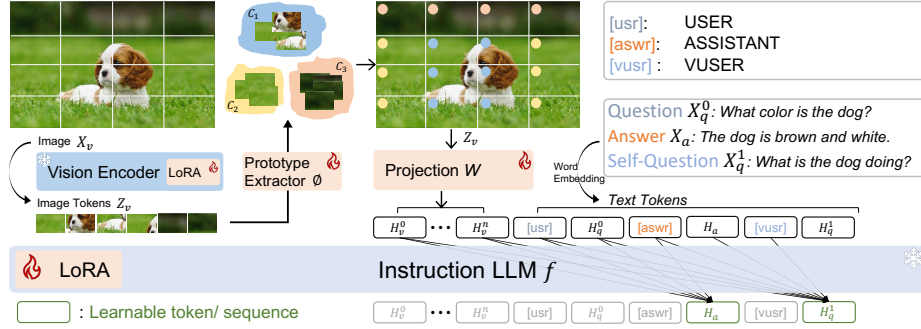


Fig. 3: Model architecture of SQ-LLaVA. We propose prototype extractor to extract visual clustering information to enhance the visual embedding encoded by the visual encoder. SQ-LLaVA defines a new token `[vusr]` as specific instruction for LLM to perform visual self-questioning. Besides question answering, SQ-LLaVA treats questioning as another training objective.

3 Method

3.1 Architecture Overview

The proposed SQ-LLaVA model (see Fig. 3) consists of four main components: 1) A pre-trained vision encoder CLIP-ViT [33] that extracts a sequence embedding of image tokens Z_v for an input image X_v ; 2) A prototype extractor $\phi(\cdot)$ learning visual clusters to enhance the original image tokens; 3) A trainable projection block $W(\cdot)$ with two linear layers to map the enhanced image tokens to the language domain tokens H_v , handling the dimension misalignment between the vision and language domain; and 4) Our LLM backbone $f(\cdot)$ implemented by the pre-trained Vicuna [57] to predict the next token upon the previous embedding sequence. Given the input question X_q and answer X_a , a word embedding matrix is used to map them to contextual embeddings H_q and H_a , and the distribution over $H_a^{(i+1)}$ can be obtained following the auto-regressive model as:

$$p_\theta(H_a^{(i+1)} | H_v, H_q, H_a^{(1:i)}) = \sigma(f(H_v, H_q, H_a^{(1:i)})), \quad (1)$$

where θ represents all the trainable parameters in our model, $\sigma(\cdot)$ is a softmax function, and $f(\cdot)$ outputs the last token embedding of the whole sequence. We denote p_θ as the prediction probability for the anticipated answer token $H_a^{(i+1)}$ at the position $i + 1$, conditioning on the input image token embedding H_v , the question token embedding H_q , and the previous answer token embeddings $H_a^{(1:i)}$. As shown in Eq. (1), the proposed SQ-LLaVA applies the language model $f(\cdot)$ to model p_θ given by H_v , H_q , and $H_a^{(1:i)}$. Existing visual instruction tuning methods [5, 21] are only able to predict H_a , yet cannot fully exploit the rich semantic clues within H_q . In this study, we propose visual self-questioning instructions to guide LLMs in capturing the visual content within questions.

System-message
 $\mathbf{X}_c^{(1)} : [\text{usr}] : \mathbf{X}_v \backslash \mathbf{X}_q^{(1)} [\text{aswr}] : \mathbf{X}_a^{(1)} < o^d >$ $\mathbf{X}_c^{(2)} : [\text{vusr}] : \mathbf{X}_q^{(2)} < o^d > [\text{aswr}] : \mathbf{X}_a^{(2)} < o^d >$
 $\mathbf{X}_c^{(3)} : [\text{usr}] : \mathbf{X}_q^{(3)} [\text{aswr}] : \mathbf{X}_a^{(3)} < o^d >$... $\mathbf{X}_c^{(P)} : [\text{vusr}] : \mathbf{X}_q^{(P)} < o^d > [\text{aswr}] : \mathbf{X}_a^{(P)} < o^d >$

Fig. 4: The input sequence used to train SQ-LLaVA. Our model is trained to predict *question*, *answer*, and *where to stop*. We use `tokens` to represent learnable tokens, where \mathbf{X}_q is the question, \mathbf{X}_a is the answer, and $< o^d >$ is the delimiter token. In SQ-LLaVA, the *System-message* = "The assistant gives helpful, detailed, and polite answers to the user's questions. Also, the assistant is a curious virtual user can ask complex questions that are relevant to the content in the image."

3.2 Visual Self-questioning Instruction

In broad real-world scenarios, proactively asking a question requires more understanding and background knowledge than answering [41]. Similarly, this work proposes visual self-questioning to encourage the LLM to discover deeper vision-language alignment and improve the overall instruction-following performance. Particularly, SQ-LLaVA treats questioning as a new learning objective beyond answering, which, to the best of our knowledge, is the first practice in the field of visual instruction tuning. While the current vision-language model can ask questions [1], such skill is still learned from question-answering through instruction tuning. However, our proposed method shows that the decoder-based LLM has the potential to learn more skills, such as how to ask questions spontaneously when a unique instruction token is given (e.g., we define `[vusr]` in our work). Furthermore, visual self-questioning can potentially improve general vision-language understanding. To be specific, as shown in Fig. 1, there are a certain amount of questions containing more meaningful image-related information than answers in the existing visual instruction data [5, 21]. Thus, we hypothesize that the vision-language understanding can be improved once the LLM learns how to predict relevant questions about a given image.

Self-Questioning Prompt Design. We provide ground-truth content for the visual self-questioning, restricting SQ-LLaVA from asking image-related questions. To this end, we leverage questions as another learnable resource and follow the regular auto-regressive training objective. As shown in Fig. 4, the training data for SQ-LLaVA is designed in a format with a pre-defined template. To be specific, the *system message* as a fixed prompt is added at the beginning of each instruction data, indicating a general job description (e.g., gives helpful answers and asks complex questions) for the LLM. Existing visual instruction tuning methods utilize the unique tokens (`[usr]`, `[aswr]`) to give the LLM a particular instruction (e.g., questioning understanding, answer prediction, etc.) and apply the delimiter token $< o^d >$ to mark the ending position. In this work, we propose a new special token `[vusr]`, indicating a specific instruction – asking questions. Combining with the delimiter, we can construct instructions for self-questioning as a new training task.

Each sample of current visual instruction data consists of one image X_v and P question-answer pairs $(X_q^{(1)}, X_a^{(1)}, \dots, X_q^{(P)}, X_a^{(P)})$. We collect one question $X_q^{(j)}$ and its answer $X_a^{(j)}$ with special tokens to construct the j^{th} turn conversation as

$$X_c^j = \begin{cases} ([usr], X_q^{(j)}, [aswr], X_a^{(j)}) & j = 1 \text{ or } j > 1, R < \delta \\ ([vusr], X_q^{(j)}, [aswr], X_a^{(j)}) & j > 1, R > \delta \end{cases}, \quad (2)$$

where $R \in [0, 1]$ is a random number and $\delta = 0.5$ is a threshold that sets the proportion of self-questioning pairs in the conversations. Finally, the full-text input X_c will be mapped to textual embedding H_c through word embedding.

SQ-LLaVA performs zero-shot questioning without any in-context knowledge or human language instruction since the only instructional prompt is a unique token `[vusr]`. After visual self-questioning on instruction data with various question formats X_q , the questions sampled by SQ-LLaVA are more diversified than GPT4-V (as shown in Fig. 5), since the LLM has learned alignment between image and questions, which is different from previous works [1, 48]. Specifically, previous general-purpose vision language models such as GPT4-V [1] can generate questions based on a given image, but it requires explicit language instruction such as “Ask complex questions that are relevant to the content in the image”. Accordingly, the quality of generated questions highly relies on prompt engineering. Also, self-instruct [48] utilizes in-context learning to prompt the LLM to ask specific questions.

3.3 Enhanced Visual Representation

Unlike previous visual instruction tuning methods, SQ-LLaVA jointly benefits from visual self-questioning and question-answering. For better visual self-questioning, we develop a prototype extractor that recognizes and groups similar patterns of visual information from the latent space. Our primary goal is to enhance visual representation through prototype learning.

Specifically, we utilize clustering to extract centroid representations of image tokens Z_v , where each cluster center is treated as a prototype, which, in return, will be distributed to each of the original image token embeddings. Our proposed prototype extractor $\phi(\cdot)$ is a lightweight design involving two parts: 1) cluster center optimization and 2) prototype information distribution. Following [14, 45], we randomly initialize $K = 256$ cluster centers C and deploy the iterative Expectation-Maximization (EM) clustering process to capture representative semantics in the latent space by

$$\begin{aligned} \text{E-step : } \mathcal{M}^{(t)} &= \sigma(q(C^{(t)}) * k(Z_v)^\top), \\ \text{M-step : } C^{(t+1)} &= \mathcal{M}^{(t)} * v(Z_v), \end{aligned} \quad (3)$$

where $\mathcal{M}^{(t)} \in [0, 1]$ denotes a soft cluster assignment map at the t^{th} iteration, $\sigma(\cdot)$ is a softmax function, and $t \in \{1, \dots, T\}$ indexes the iteration of EM step with $T = 2$ in this work. Three trainable linear layers q , k , and v are used in (3.3),

where $q(\cdot)$ projects the prototype C to a query vector, and $k(\cdot)$ and $v(\cdot)$ project H_v into key and value vectors, followed by a normalization layer, respectively. The prototype extractor iteratively updates cluster map \mathcal{M} and centers C .

After the prototype extraction, we train a linear layer $z(\cdot)$ to adaptively map the visual cluster information to the raw image embedding Z_v . For the i^{th} token embedding $Z_v^{(i)}$, we update it as

$$Z_v^{(i)} = Z_v^{(i)} + z\left(\frac{1}{K} \sum_{j=1}^K S_c(C_j, Z_v^{(i)}) \times C_j\right), \quad (4)$$

where $S_c(\cdot, \cdot)$ is a normalized cosine similarity function. The weighted sum over prototypes in Eq. (4) emerges as an indispensable step for contextual understanding from image tokens, recognizing and grouping similar patterns and semantics. It clusters image tokens as prototypes that display homogeneity in semantics, such as “grass” and “dog”. The prototypes can describe the intrinsic semantic meanings by aggregating entities that exhibit shared attributes. Finally, we map the image sequence embedding Z_v to language domain H_v with a two-layer linear projector $W(\cdot)$.

3.4 Model Training

Stage1: Pre-training for Vision-Language Alignment. Unlike text-only LLMs, the vision-language model also fine-tunes LLM using image tokens as input (see Fig. 3). Therefore, the pre-training stage aims to optimize the LLM by explicitly executing the visual instruction. The proposed SQ-LLaVA adopts Vicuna [57] as its instruction LLM, pre-training on massive text corpora to predict the next text token given the previous context, not only containing text but also visual instructions. To achieve this, we organize the pre-training data as $D_{PT} = \{[X_v^{(1)}, X_a^{(1)}], \dots, [X_v^{(N)}, X_a^{(N)}]\}$, where N is the total number of training samples, and each sample has an image and its related descriptions. Each image and text input pair will be mapped to sequence embeddings (H_v and H_a) as elaborated in Section 3.1. During pre-training, we freeze the vision encoder and LLM and mainly train the prototype extractor ϕ and the vision-to-language projector W . The pre-training goal is to maximize the probability of the predicted image description H_a given an image H_v . When training a visual instructional LLM, we follow the negative log-likelihood objective function as

$$\sum_{v,a \in D_{PT}} -\log p_{\theta}(H_a | H_v) = \sum_{v,a \in D_{PT}} \sum_{i=1}^L -\log p_{\theta}(H_a^{(i+1)} | H_v, H_a^{(1:i)}), \quad (5)$$

where L denotes the sequence length of answer tokens in H_a , θ is the total trainable parameter of ϕ and W , $p(H_a | H_v)$ can be computed by Eq. (1), and $H_a^{(1:i)}$ represents all the answer tokens before the current prediction $H_a^{(i+1)}$.

Stage2: Fine-tuning. Existing methods, such as LLaVA [5, 22], mainly update the vision-to-language projector (usually a couple of linear layers) and the

language model during fine-tuning. Nevertheless, the projector might be too weak to capture the relationship between the image and the questions. Following the previous multi-modal understanding method [33], we unfreeze the vision encoder and LLM during fine-tuning for a joint optimization further to eliminate the gap between the vision and language domain.

To mitigate the heavy computational overhead, we take advantage of LoRA [12] as a lightweight tuning option that can achieve similar (even better) performance to fully fine-tuning when training large models on a relatively small dataset. We add LoRA in both the vision encoder and LLM. Thus, the learnable parameters θ of the proposed SQ-LLaVA during fine-tuning represent a combination of all the parameters of LLM-LoRA, ViT-LoRA, prototype extractor ϕ , and the vision-to-language projector W . Given the instruction tuning data $D_{IT} = \{[X_v^{(1)}, X_c^{(1)}], \dots, [X_v^{(N)}, X_c^{(N)}]\}$, we take the conversational data X_c and the image X_v as input, mapping them to sequence embedding (H_c and H_v) as elaborated in Section 3.1, and minimize the negative log-likelihood loss for the *self-questioning* and *answering* tasks as follows

$$\text{Self-questioning : } \sum_{v,c \in D_{IT}} -\log p_{\theta}(H_q^{(j+1)} \mid H_v, H_c^{(1:j)}), \quad (6)$$

$$\text{Answering : } \sum_{v,c \in D_{IT}} -\log p_{\theta}(H_a^{(j+1)} \mid H_v, H_c^{(1:j)}, H_q^{(j+1)}), \quad (7)$$

where $j \in \{1, \dots, P\}$, indicating the index of question or answer within the conversational data $X_c^{(*)}$. Notably, previous works [5, 21, 58] only involve answering tasks, but we introduce visual self-questioning as an additional training task for visual instruction tuning. Eventually, SQ-LLaVA, as a vision-language assistant, not only executes human instructions by optimizing the objective function in Eq. (7) but can raise questions out of the given image after optimizing the Eq. (6). This capability potentially yields more diverse question-answer guidance and enhances multi-modal understanding.

4 Experiments

4.1 Experimental Setting

Dataset. Our work uses the open-source visual instruction dataset provided by LLaVA [21] and ShareGPT4V [5] for training. Each dataset has large-scale image-text paired data for pre-training and instruction-following data for fine-tuning. Specifically, the instruction data proposed by LLaVA is a comprehensive mixture of COCO, GQA [13], OCR-VQA [27], TextVQA [39], VisualGenome [17], RefCOCO [15], and image from InstructBLIP [8] involves multiple reasoning, spatial understanding, multi-step inference, optical character recognition, and grounding of visual concepts to language. The ShareGPT4V dataset consists of the same mixture image as LLaVA but enrolls more images from other datasets such as SAM [16] and WebData [29, 36].

We evaluate our model on general vision-language understanding tasks. Specifically, we use ten visual oriented question answering benchmarks: VQA^{v2} [10]; GQA [13]; VizWiz [11]; SQA^I [25]; ScienceQA-IMG; VQA^T: TextVQA [39]; POPE [20]; MM-Vet [53]; LLaVA^W [22]; LLaVA (in the wild); MMB: MMBenchmark [23]; MMB^{CN}: MMBench-Chinese [23]. We report the prediction accuracy on all the benchmarks. We also evaluate our model for visual information discovery through captioning. To be specific, we employ testing images from four datasets, *i.e.*, COCO [6], Flickr30K [52], Conceptual [37], and Nocaps [2]. Following [9], we evaluate the proposed method with regular image captioning metrics, *e.g.*, BLEU [32] and CIDEr [46]. For the open-world methods LLaVA-v1.5 [21], ShareGPT4V [5] and the proposed SQ-LLaVA, we utilize greedy search for caption generation with a prompt of “Provide a brief description of the given image, your answer should be in one sentence.” The generated caption is used to evaluate the performance of image captioning.

4.2 Implementation

We adopt Vicuna [57] as the pre-trained language generative model and CLIP-ViT [33] as the vision encoder. We pre-train the prototype extractor and the vision-to-language projector using AdamW [24] optimizer with a learning rate of 2×10^{-3} and a constant scheduler for one epoch. Following previous work [21], we keep the global batch size as 256 for pre-training and 128 for fine-tuning. During fine-tuning, we insert LoRA [12] with $rank = 128$ and $\alpha = 256$ into the language model (LLM-LoRA) and LoRA with $rank = 32$ and $\alpha = 64$ into the vision encoder (ViT-LoRA). We optimize LoRA modules, the prototype extractor, and vision-to-language projector for one epoch. We set the learning rate to 2×10^{-4} for LoRA, and 2×10^{-5} for the other layers. All the weights of the pre-trained language model and vision encoder remain fixed during fine-tuning.

4.3 Zero-shot Multilingual Capability

We evaluate SQ-LLaVA on ten benchmarks, covering a range of academic Visual Question Answering (VQA) tasks and recent instruction tuning tasks designed for large vision language models. The academic VQA includes VQA-v2 [10] and VizWiz [11]. GQA [13] is a fine-grained real-world visual reasoning and question-answering benchmark. ScienceQA [25] is a benchmark with rich subjects (natural science, language science, and social science). TextVQA [39] requires the model to recognize the texts in the image. LLaVA (in the wild) and MM-Vet [53] use GPT4 to assess the capability of the models for testing. With manually designed questions, MM-Bench and MMBench-CN [23] evaluate the model’s vision-related reasoning and perception for English and Chinese, respectively. POPE [20] is a benchmark for evaluating object hallucination [34].

In Table 1, we quantitatively compare between SQ-LLaVA and existing models. SQ-LLaVA-7B and SQ-LLaVA-13B trained on two instruction datasets [5, 21] outperform previous methods in six out of ten visual instruction tuning tasks. To be specific, SQ-LLaVA-7B achieves 17.2% improvement over LLaVA-v1.5-7B

Table 1: Comparison with state-of-the-art methods on ten benchmarks. After training on the same instruction data, SQ-LLaVA trained on [21] and SQ-LLaVA* trained on [5] surpass their baseline model LLaVA-v1.5 and ShareGPT4V on 9 out of 10 and 6 out of 10 benchmarks in the 7B scale, and 8 out of 10 and 6 out of 10 in the 13B scale. The best results are **bold** and the second-best results are underlined.

LLM Model		VQA ^{v2}	GQA	VizWiz	SQA ^I	VQA ^T	POPE	MM-Vet	LLaVA ^W	MMB	MMB ^{CN}
7B	InstructBLIP [8]	-	49.2	34.5	60.5	50.1	-	26.2	60.9	36.0	23.7
	Qwen-VL [3]	78.8	59.3	35.2	67.1	63.8	-	-	-	38.2	7.4
	Qwen-VL-chat [3]	78.2	57.5	38.9	68.2	<u>61.5</u>	-	-	-	60.6	56.7
	LLaVA-v1.5 [21]	78.5	62.0	50.0	66.8	58.2	85.9	30.5	63.4	64.3	58.3
	ShareGPT4V [5]	80.6	<u>63.3</u>	57.2	68.4	60.4	86.8	37.6	<u>72.6</u>	68.8	62.2
	SQ-LLaVA	79.2	62.8	54.0	<u>68.9</u>	58.6	87.7	32.5	66.3	66.2	58.1
	SQ-LLaVA*	<u>80.3</u>	63.7	<u>55.3</u>	70.5	60.5	<u>87.2</u>	37.6	74.3	<u>66.6</u>	<u>60.0</u>
13B	InstructBLIP [8]	-	49.5	33.4	63.1	50.7	78.9	25.6	58.2	-	-
	LLaVA-v1.5 [21]	80.0	63.3	53.6	71.6	58.2	85.9	35.4	70.7	67.7	<u>63.6</u>
	ShareGPT4V [5]	<u>81.0</u>	<u>64.8</u>	<u>55.6</u>	71.2	62.2	-	43.1	<u>79.9</u>	<u>68.5</u>	63.7
	SQ-LLaVA	80.1	63.6	54.6	69.8	60.2	87.7	35.5	74.6	68.7	62.0
	SQ-LLaVA*	81.3	65.0	58.2	<u>71.5</u>	<u>61.9</u>	<u>87.4</u>	<u>39.7</u>	80.7	<u>68.5</u>	62.5

on the LLaVA (in the wild) benchmark, indicating the superior capabilities of our model in tasks such as detailed description and complex reasoning. Also, SQ-LLaVA-7B improves over previous methods on ScienceQA, indicating that our model excels in understanding and reasoning over scientific content and can effectively handle multi-modal information. The improvement in ScienceQA suggests strong capabilities in multi-hop reasoning, comprehension of complex scientific concepts, and the ability to utilize context and explanations to derive correct answers. SQ-LLaVA-7B has a steady improvement over LLaVA-v1.5-7B and ShareGPT4V-7B on the POPE benchmark, and the 2% and 1% improvement indicates that our proposed method has better reliability and trustworthiness since POPE is a task designed to evaluate the phenomenon of object hallucination [20, 34]. In the bottom section of Table 1, the proposed SQ-LLaVA-13B surpasses previous works in six out of ten benchmarks, indicating the scalability of our method on larger LLM. Notably, the performance inconsistency on some datasets might be due to the unsupervised prototype extractor (lacking pixel-wise guidance) in our model. To mitigate this issue, we could leverage pseudo object masks (e.g., given by the pre-trained segment anything) in learning prototypes. Despite this limitation, all the improvements are achieved with significantly fewer trainable parameters compared to other methods [3, 5, 21].

4.4 Visual Information Discovery

In this experiment, we showcase the diversity and reliability of the proposed SQ-LLaVA through various qualitative applications, including detailed image description, visual information summary, and visual self-questioning. We also present quantitative results on the task of image captioning.

Abilities of SQ-LLaVA Through Qualitative Samples. SQ-LLaVA exhibits numerous advanced capabilities compared to traditional vision-language models.

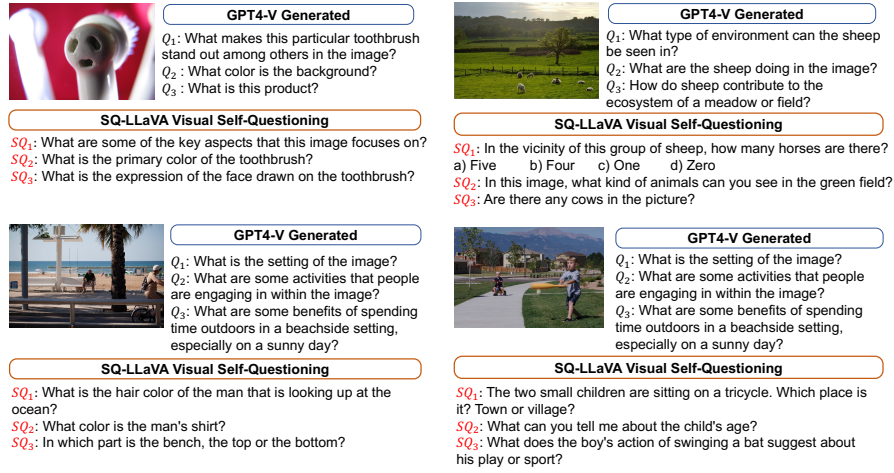


Fig. 5: Visual self-questioning of SQ-LLaVA-7B. Comparing to the question data provided by GPT4-V (data collected by LLaVA-v1.5 [21]), SQ-LLaVA can generate questions with higher diversity *i.e.* multiple choice and tricky questions.

Table 2: Comparison of image captioning on four datasets. SQ-LLaVA is trained on data collected by [21]. SQ-LLaVA* is trained on data collected by [5].

Model	Flickr30k		Nocaps ^{out}		Conceptual	
	B@4	CIDEr	B@4	CIDEr	B@4	CIDEr
ClipCap [28]	17.21	41.65	20.32	51.74	1.47	23.74
DiscriTune [9]	18.48	44.78	24.10	57.06	1.71	28.01
LLaVA-v1.5 [21]	28.67	81.27	35.78	103.56	2.79	39.20
ShareGPT4V [5]	31.00	86.17	37.19	107.45	2.78	37.86
SQ-LLaVA	29.88	83.51	36.21	105.42	2.90	39.49
SQ-LLaVA*	31.49	83.14	37.20	107.42	2.91	41.24

(a) Zero-shot image captioning

Model	Zero-shot		Fine-tune	
	B@4	CIDEr	B@4	CIDEr
ClipCap	8.50	37.03	32.60	108.55
DiscriTune	13.99	53.20	32.31	105.40
eP-ALM [38]	29.47	97.22	33.35	111.63
MAPL [26]	12.30	54.30	36.45	125.20
LLaVA-v1.5	29.96	111.46	-	-
SQ-LLaVA	29.85	110.77	40.76	136.78

(b) Image captioning on COCO

Notably, SQ-LLaVA effectively mitigates object hallucination [20, 34], resulting in predictions that are more trustworthy. Additionally, SQ-LLaVA is capable of generating diverse and meaningful questions about a given image without requiring human textual instructions. As illustrated in Fig. 5, the generated questions include multiple-choice, reasoning, binary choice questions, and the content within each question, all of which are either highly correlated or tricky to the given image, such as “What is the hair color of the man that is looking up at the ocean?” from the third image and “Are there cows in the picture?” from the second image. These examples highlight SQ-LLaVA’s ability to produce questions with high diversity and quality.

The detailed image description example in Fig. 6 demonstrates that SQ-LLaVA can generate visual information with concrete concepts within the image, such as “Hyundai”, “small white dog”, “Nikky Stephen”. In contrast, LLaVA-v1.5 can

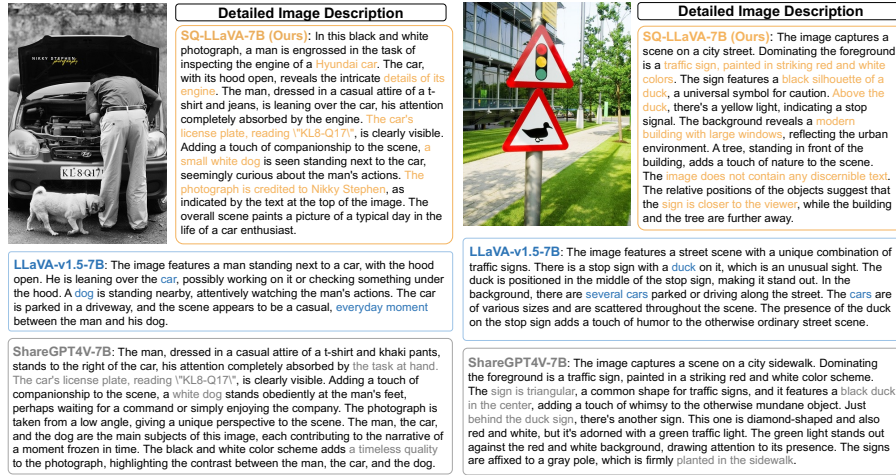


Fig. 6: A qualitative evaluation of detailed image descriptions from three models. We highlight the words and sentences that represent how each model describes the main object in the image.

only describe the image from a general concept, such as “car”, “dog”, or “everyday moment”. Even though ShareGPT4V can generate detailed descriptions, it still suffers from the issue of object hallucination, such as “perhaps waiting” and “behind the duck sign”. By observation, SQ-LLaVA can describe the image with less unintended text, yielding higher reliability.

Quantitative Analysis. SQ-LLaVA serves as a general-purpose vision-language model and enables zero-shot image captioning. As indicated in Table 2a, SQ-LLaVA achieves 73% and 66% averaged improvement over ClipCap and DiscriTune on all datasets, indicating the effectiveness of visual instruction tuning. Compared with the baseline model LLaVA-V1.5 [21], SQ-LLaVA achieves 2% averaged improvement on all datasets, indicating the effectiveness of visual self-questioning. Also, we find SQ-LLaVA* surpasses the baseline model ShareGPT4V [5] on Nocaps^{out} and Conceptual dataset, demonstrating the adaptability of the proposed method on unseen testing images from new domains. Moreover, as shown in Table 2b, SQ-LLaVA can be easily adapted to COCO captioning via instruction tuning on short descriptions.

4.5 Ablation Study

In Table 3, we conduct experiments with different architecture designs and training strategies on five question-answering benchmarks. For a fair comparison, we train the baseline models on our local machine with the same training recipe of LLaVA-LoRA [21]. Specifically, we present our full model and three ablated models by removing one component each time. We adopt the dataset [21] with 558k for pre-training (PT) and 665k for fine-tuning (FT). As compared, self-

Table 3: Ablation study of training strategy on visual instruction tasks. All models are in 7B scale with three components of ViT-LoRA (V-LoRA), self-questioning (SQ), and prototype extractor (Proto). We provide experiments on two instruction datasets [5, 21] 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	✗	✗	✗	49.4	68.4	58.2	86.5	67.1	65.9
		✓	✗	✓	52.4	67.9	58.6	87.7	65.6	66.4
		✓	✓	✗	52.6	68.4	57.8	88.2	67.3	66.9
		✗	✓	✓	53.4	69.3	58.1	87.9	67.9	67.3
		✓	✓	✓	54.0	68.9	58.6	87.7	68.1	67.5
1200K	700K	✗	✗	✗	51.5	68.9	58.9	86.8	72.1	67.6
		✓	✗	✓	54.0	68.9	60.2	87.2	71.6	68.4
		✓	✓	✗	55.4	69.2	59.5	86.8	77.3	69.6
		✗	✓	✓	54.2	70.3	60.5	87.5	72.7	69.0
		✓	✓	✓	55.3	70.5	60.5	87.2	74.3	69.6

questioning (SQ) brings a consistent performance boost on all the five benchmarks, indicating the effectiveness of visual self-questioning on improving visual language understanding. Besides, we introduce the prototype extractor (Proto) to enhance visual representation, achieving 0.9% improvement in average accuracy among five benchmark. With all three components incorporated, we observe a 2.4% improvement in average accuracy.

As shown by the bottom block of Table 3, we conduct experiments with the same ablation settings but with a larger scale of the visual instruction data [5] (*i.e.*, for both PT and IT). Overall, SQ-LLaVA achieves 2.4% improvement over the baseline model after training on the smaller dataset and achieves 3.0% improvement after training on the larger dataset.

5 Conclusions

This work has introduced SQ-LLaVA, a new visual instruction tuning method that enhances general-purpose vision-language understanding and image-oriented question answering through visual self-questioning. Our experiments demonstrate that SQ-LLaVA achieves superior performance with fewer training parameters and instructional data. We have also conducted a comprehensive study on visual discovery/reasoning tasks and found that SQ-LLaVA generalizes well to a wide range of unseen tasks and outperforms several state-of-the-art methods. Qualitative assessments show that SQ-LLaVA strengthens visual representation and domain alignment, effectively reducing object hallucination and improving the semantic interpretation of images. Our findings highlight the potential of visual self-questioning as a powerful training strategy for the visual instruction tuning framework, paving the way for realizing more efficient and effective large vision-language models. Particularly, SQ-LLaVA frames questioning as an intrinsic goal of tuning LLMs, encouraging the exploration of the model’s curiosity (the ability to ask questions proactively) in solving complex problems.

References

1. Gpt-4v(ision) system card (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: ICCV (2019)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. In: ArXiv (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS (2020)
5. 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)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. In: ArXiv (2015)
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N.M., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B.C., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., García, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Díaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K.S., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways. In: J. Mach. Learn. Res. (2022)
8. 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)
9. Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N.C., Franzon, F., Baroni, M.: Cross-domain image captioning with discriminative finetuning. In: CVPR (2023)
10. 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: CVPR (2017)
11. Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
13. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
14. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. In: Information Sciences (2023)

15. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. In: ICCV (2023)
17. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: IJCV (2016)
18. Li, J., Pan, K., Ge, Z., Gao, M., Zhang, H., Ji, W., Zhang, W., Chua, T.S., Tang, S., Zhuang, Y.: Empowering vision-language models to follow interleaved vision-language instructions. In: ICLR (2024)
19. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)
20. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: ArXiv (2023)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
23. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: ArXiv (2023)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. 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. In: NeurIPS (2022)
26. Mañas, O., López, P.R., Ahmadi, S., Nematzadeh, A., Goyal, Y., Agrawal, A.: Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In: EACL (2023)
27. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019)
28. Mokady, R.: Clipcap: Clip prefix for image captioning. In: ArXiv (2021)
29. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS (2011)
30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askill, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askill, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: ACL (2002)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askill, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)

34. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018)
35. Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A., Dey, M., Bari, M.S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N.V., Datta, D., Chang, J.D., Jiang, M.T.J., Wang, H., Manica, M., Shen, S., Yong, Z.X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Févry, T., Fries, J.A., Teehan, R., Biderman, S., Gao, L., Bers, T., Wolf, T., Rush, A.M.: Multitask prompted training enables zero-shot task generalization. In: ICLR (2022)
36. 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. In: NeurIPS Workshop (2021)
37. 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) (2018)
38. Shukor, M., Dancette, C., Cord, M.: ep-alm: Efficient perceptual augmentation of language models. In: ICCV (2023)
39. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
40. Sun, G., Bai, Y., Yang, X., Fang, Y., Fu, Y., Tao, Z.: Aligning out-of-distribution web images and caption semantics via evidential learning. In: Proceedings of the ACM on Web Conference 2024 (2024)
41. Tofade, T., Elsner, J., Haines, S.: Best practice strategies for effective use of questions as a teaching tool. In: American journal of pharmaceutical education (2013)
42. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. In: ArXiv (2023)
43. Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D.M., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A.S., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I.M., Korenev, A.V., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. In: ArXiv (2023)
44. Triantafillou, E., Zhu, T.L., Dumoulin, V., Lamblin, P., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., Larochelle, H.: Meta-dataset: A dataset of datasets for learning to learn from few examples. In: ICLR (2019)
45. Vattani, A.: K-means requires exponentially many iterations even in the plane. In: Annual Symposium on Computational Geometry (2009)
46. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
47. Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Text is mass: Modeling as stochastic embedding for text-video retrieval. In: CVPR (2024)

48. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. In: ACL (2022)
49. Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A.S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H.G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., Pal, K.K., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P.R., Verma, P., Puri, R.S., Karia, R., Sampat, S.K., Doshi, S., Mishra, S.D., Reddy, S., Patro, S., Dixit, T., Shen, X., Baral, C., Choi, Y., Smith, N.A., Hajishirzi, H., Khashabi, D.: Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In: EMNLP (2022)
50. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: ICLR (2022)
51. Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Jiang, D.: Wizardlm: Empowering large language models to follow complex instructions. In: ICLR (2024)
52. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In: TACL (2014)
53. 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. In: ArXiv (2023)
54. 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. In: ICLR (2024)
55. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: Llavav: Enhanced visual instruction tuning for text-rich image understanding. In: ArXiv (2023)
56. Zhao, B., Wu, B., Huang, T.: Svit: Scaling up visual instruction tuning. In: ArXiv (2023)
57. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: NeurIPS (2023)
58. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)