




Questioning, Answering, and Captioning for Zero-Shot Detailed Image Caption

Duc-Tuan Luu^{1,2,3,4} , Viet-Tuan Le⁵ , and Duc Minh Vo⁶ 

¹ University of Information Technology, VNU-HCM, Vietnam

² University of Science, VNU-HCM, Vietnam

³ John von Neumann Institute, VNU-HCM, Vietnam

⁴ Vietnam National University, Ho Chi Minh City, Vietnam

⁵ Ho Chi Minh City Open University, Vietnam

⁶ The University of Tokyo, Japan

tuanld@uit.edu.vn, tuan.lv@ou.edu.vn, vmduc@nlab.ci.i.u-tokyo.ac.jp

Abstract. End-to-end pre-trained large vision language models (VLMs) have made unprecedented progress in image captioning. Nonetheless, they struggle to generate detailed captions, which necessitate the models capturing spatial relations, counting, text rendering, world knowledge, and other presenting or not presenting aspects of the image. To overcome their inadequacies, we present a *Question – Answer – Caption* methodology, named QAC, that performs questioning and answering on many aspects of the given image, followed by captions based on the responses. Specifically, we use ChatGPT to produce a set of questions about the images' content. The questions are then answered using a pre-trained VLM. After gathering all answers, we prompt the pre-trained VLM to generate descriptive captions in a zero-shot setting. Our approach is plug-and-play and can be easily applied on any pre-trained VLM. We implement QAC on InstructBLIP and LLaVA, demonstrating comparable performance to fine-tuned models on a challenging DOCCI dataset.

1 Introduction

The field of image captioning has witnessed remarkable advancements with the emergence of VLMs [2, 32, 55, 57, 60]. These models have demonstrated impressive capabilities in generating detailed and informative captions based on visual input, making VLMs increasingly important across many downstream vision-language tasks. Usually, they are designed in a single-step paradigm, which returns a final caption directly from an image. This approach, however, hurts the performance when it comes to zero-shot reasoning [59] or visual entailment [53] tasks, which require in-depth captions from multi-step reasoning. Particularly, existing VLMs often struggle to capture the intricacies of complex visual scenes, leading to captions that lack the specificity and richness necessary for accurate and informative descriptions. Hence, this drawback hinders VLMs applicability in various domains, such as content creation, image search, accessibility, etc.

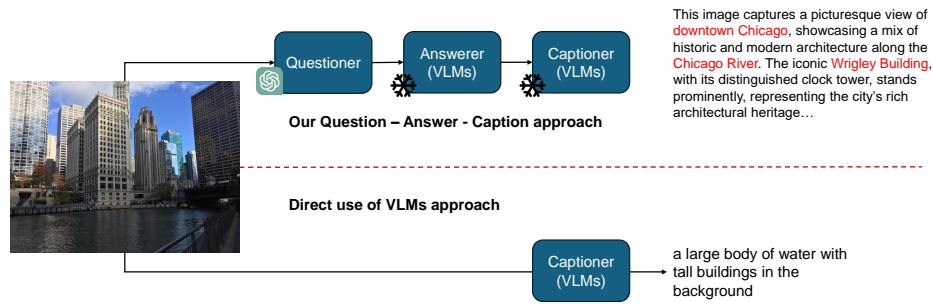


Fig. 1: Compared to traditional single-step image captioning VLMs, our multi-step QAC approach provides more insights and generates a higher level of descriptive captions, mimicking the process of human reasoning and understanding.

In contrast, with a random photo, humans can provide a comprehensive fine-grained caption with ease and proficiency. Humans would intuitively break down the general image captioning task into different questions related to various aspects of the image. Moreover, the questions may even related one after another, acting like a human reasoning process from cumulative data. After answering all the questions, we can combine all the information and produce an insightful and descriptive caption, containing a deep understanding of multiple aspects, including objects, attributes, views, scenes, spatial relationships, text rendering, and world knowledge.

Inspired by this human nature behaviour, we propose a novel method that leverages the power of VLMs [1, 4, 12, 47, 49] to enhance the detail and accuracy of generated captions. Our method is reinforced by the observation that large language models excel at understanding and generating textual data, while deep learning visual models [10, 23, 26, 27, 36, 39] are adept at processing and understanding visual information. We aim to create a robust and effective image captioning system by combining these strengths in a simple yet effective *Question - Answer - Caption* (QAC) paradigm. Figure 1 demonstrates the workflow of our proposed method, which contains three main components: the *Questioner*, *Answerer*, and *Captioner*. Based on the visual content of the input image, the *Questioner* generates a set of questions with different perspectives. The corresponding descriptive answers are obtained by utilizing the *Answerer*. Finally, the *Captioner* is responsible for synthesizing a fine-grained caption by analyzing and combining the information from the set of answers.

We carefully evaluate our proposed QAC image captioning method on the manually annotated image-caption pairs DOCCI dataset [33]. We implement our approach on the top of two open-source VLMs, including InstructBLIP (Vicuna-7B) [11, 13] and LLaVA-1.5 7B [32] in a zero-shot manner, obtaining a marginal improvement on conventional metrics. We also notice that our generated captions contain information which does not present in the image such as world knowledge (words highlighted in red in Figure 1, 2 and 7), revealing that *Question - Answer*

– *Caption* is a good guidance for the image captioning task. Our contributions are summarized as follows:

- We introduce a straightforward yet effective approach called QAC, which is a *Question – Answer – Caption* framework designed to enhance the generation of detailed image captions in a zero-shot setting.
- Our method is designed to be plug-and-play, making it fully compatible with any pre-trained VLMs without requiring any additional re-training or fine-tuning.

2 Related Work

2.1 Multi-Modality Models

Recent developments in multi-modality models [14, 26, 31, 32, 60] have notably enhanced the capabilities of comprehension and reasoning. These models often leverage the alignment of pre-trained large vision models [10, 23, 39, 45] with large language models [4, 11, 12, 49]. Early works, such as BLIP [26, 27], LLaVA [24, 31, 32], and the Qwen series [31, 32] bridged the modality gap through resampling or MLP projectors, demonstrating promising results. The Emu series [44, 46] portrayed exceptional in-context learning ability for multi-modal content. Lately, there has been a growing trend towards developing high-resolution capability models in this research topic. Models like Monkey [29] and CogAgent [20] have adopted strategies to handle large images effectively, either by dividing them into patches or using separate low-resolution and high-resolution encoders. LLaVA-NEXT [24] and LLaVA-UHD [54] have introduced dynamic image aspect ratios, image partitioning and slicing techniques to capture more visual details. Moreover, Scaling on Scales [42] has demonstrated the ability to extract multi-scale features directly through image wrapping and rescaling without requiring an increase in image tokens.

2.2 Vision-Language Datasets

Early vision-language datasets were manually constructed using human annotations, such as Flickr30k [56] and COCO [30]. While these datasets offered high-quality annotations, they were limited in size and length. To address this, researchers turned to web-crawled datasets like YFCC100M [48] or RedCaps [15], which offered larger scales but faced challenges in terms of annotation quality. Many of these captions were only loosely related or unrelated to the corresponding images, impacting overall performance. To mitigate this issue, automatic filtering procedures were introduced to select higher-quality data samples, as seen in Localized Narratives [34] and Conceptual Captions [5, 41]. These efforts have continued to improve, resulting in billion-scale datasets like LAION-5B [40] and LAION-CAT [35], playing a crucial role in advancing vision-language

pre-training. LaCLIP [16] and CapsFusion [58] have leveraged LLMs for caption rewriting and consolidation. Moreover, recent studies have turned to GPT-4V or human-in-the-loop strategies to acquire detailed description datasets. ShareGPT4V [8] and ALLaVA [7] have generated large-scale synthetic datasets with detailed captions using GPT-4V. GLaMM [38] and all-seeing projects [51, 52] have focused on region-level vision recognition and conversation generation. ImageInWords [17] and DOCCI [33] have introduced human-in-the-loop annotation frameworks for fine-grained detailed captions.

2.3 Dense Image Captioning

Dense image captioning has gained significant attention in recent years as a means to generate multiple descriptive captions for various objects and regions within a single image. This approach diverges from traditional image captioning, which typically produces a single sentence summarizing the entire image. Early works in this field, such as DenseCap [22], laid the groundwork by integrating visual features with textual descriptions, demonstrating the potential for more detailed image analysis. Subsequently, Anderson et al. [3] introduced the bottom-up and top-down attention mechanism, which effectively combined object detection with caption generation, allowing models to focus on salient areas of an image and generate context-aware descriptions. However, these pioneer approaches could not exploit all of the complementary nature of local and global visual cues. Hence, the generated captions are often concise, neglecting details about the intricate relation between objects. To alleviate this problem, recent studies [6, 18, 19, 21] have focused on combining both LLMs and VLMs to boost the ability to generate high-fidelity dense captions. GBC [21] was proposed as a new vision-language data format that captions images with a graph-based structure akin to scene graphs while retaining the flexibility and intuitiveness of plain text description. VCB [18] introduced a blended mechanism that holistically captures various perspectives of the image while remaining anchored in human annotations. VFC [19] initiated a verification step after generating caption proposals by using tools such as object detection and visual question answering (VQA) models. This approach mitigates the challenge of hallucination in long captions. Furthermore, high-quality datasets (DenseFusion-1M [28], Pixel-Prose [43], DCI [50], DOCCI [33]) are also proposed, having precise and reliable captions that can capture all of the aspects of the image (objects, attributes, spatial relations, scene, etc.). These datasets can benefit not only the image captioning task but also various vision-language tasks in general.

3 Question – Answer – Caption Method

We propose a plug-and-play novel method called *Question – Answer – Caption* (QAC) to enhance detailed image captioning in a zero-shot setting, leveraging pre-trained VLMs. Our method consists of three effective components that operate consecutively, namely the *Questioner*, *Answerer* and *Captioner*. The

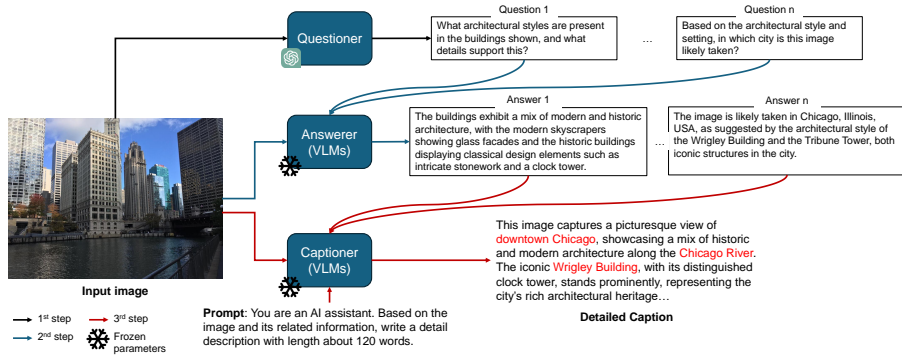


Fig. 2: The overall pipeline of our QAC approach. It consists of three sequential steps. The first step (*Questioner*) uses ChatGPT to generate a set of questions regarding the input image. The second step (*Answerer*) prompts a pre-trained VLM to answer the questions. The final step (*Captioner*) share the same VLM with a modified prompt to obtain the final detailed caption. We highlight the **world knowledge** extracted from the image that is successfully included in the caption.

core idea behind QAC is to structure the image caption generation process as a sequence of questions and answers. Instead of solely relying on the model’s direct image-captioning capabilities, we first generate a set of n relevant questions $\mathcal{Q} = \{q_1, q_2, \dots, q_n\}$ about the image I and then use the pre-trained VLM to answer these questions, obtaining a set of n answers $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$. These answers are then used to construct a more comprehensive and enriched image caption c . Figure 2 illustrates a detailed example of each component in our proposed QAC approach for image captioning.

3.1 Questioner

```

###Human: <Image><I></Image> Ask 10 questions about this image. The
question focuses on object and attributes, spatial relationships,
text rendering, world knowledge, view/scene.
###ChatGPT:
    
```

Fig. 3: ChatGPT prompt guidance for *Questioner*.

As mentioned earlier, the *Questioner* is responsible for creating a set of questions \mathcal{Q} related to the visual information in the image. To ensure the generated captions capture multiple aspects of the image content, the *Questioner* must be capable of asking about a diverse range of characteristics, such as objects, attributes, spatial relationships, text rendering, world knowledge, and details about the view and scene. This comprehensive questioning ensures that all relevant information in the image is represented.

To achieve this level of detail, we select a powerful closed-source model as the *Questioner*, since open-source models do not produce satisfactory results, as discussed in our later analysis. Because we do not have direct access to this model on our machine, we upload the image I along with a prompt (shown in Fig. 3) through the API. Additionally, we specify the number n of questions q_i ($i = 1, \dots, n$) to be generated, aiming for comprehensive coverage. As a result, we obtain a set \mathcal{Q} consisting of n questions. Due to API costs, we conducted the questioning process only once without verifying whether all questions addressed every desired aspect.

3.2 Answerer

```
###Human: <Image><I></Image> <Questions><q1, q2, ..., qn></Questions>
You are AI assistant with immersive knowledge about the world.
Based on that, answer the question related to the image.
###VLMs:
```

Fig. 4: Prompt guidance for *Answerer*.

The second step of our proposed QAC method is the *Answerer*. This module is required to answer each question in \mathcal{Q} generated from the *Questioner* based on the image content. Therefore, we employ a pre-trained VLM to serve as the *Answerer*.

Specifically, we first extract the image features \mathbf{I} from the given image I using the image encoder in VLM. Then, we feed image features \mathbf{I} along with all questions q_i in \mathcal{Q} , achieving corresponding answers $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$. The answer prompt is illustrated in Figure 4, describing the VLM as an AI assistant with vast knowledge about the world. In fact, the content of this prompt can be manually changed to produce better answers. Moreover, these answers may contain details about world knowledge (words highlighted in red in Figure 1, 2 and 7), giving people more insights about what information presents in the photo. After this step, a set \mathcal{A} of n answers is ready to go through the *Captioner* to generate the final caption.

3.3 Captioner

```
###Human: <Image><I></Image> <Info><a1, a2, ..., an></Info> Given the
image and relevant information, describe this image in detail.
###VLMs:
```

Fig. 5: Prompt guidance for *Captioner*.

Captioner is the final stage of our proposed QAC approach. This step shares the same VLM with the *Answerer* but integrates with a different guidance prompt. The image features \mathbf{I} from the previous step continue being the input, along with the relevant information from the set of answers \mathcal{A} . Figure 5 shows how we modify the guidance prompt to achieve the fine-grained caption c from the input image I . Similar to the previous step, the guidance caption prompt could also be altered to get a better image captioning result.

4 Experimental Results

4.1 Experimental Settings

Implementation. For the *Questioner*, we employ ChatGPT 4o to generate $n = 10$ questions for each image. Since the *Answerer* and *Captioner* in our QAC are flexible, we select two different methods, InstructBLIP [13] and LLaVA-1.5 7B [32], for each version. They are denoted as QAC-InstructBLIP and QAC-LLaVa, respectively. We note that all versions of our QAC are zero-shot settings. We tested our method on a single A100 80G GPU.

Evaluation dataset. We evaluate QAC on DOCCI dataset [33] using its test set in a zero-shot manner without any training or finetuning. It contains 5,000 images with one caption for each. On average, each caption has 7 sentences with a length of 135 words. The caption covers multiple aspects, including objects, attributes, views, scenes, spatial relationships, text rendering, and world knowledge.

Compared methods. We mainly compare our method with InstructBLIP [13] (Vicuna-7B), LLaVA-1.5 7B [32] since they are able to generate detailed captions. For other SOTAs on the image captioning task, such as MiniGPT4 [60], SmallCap [37], EVCap [25], since they focus on generating short captions, we only show their generated captions.

4.2 Case Study

In the case study, illustrated by Figure 6, our method is demonstrated step-by-step. The process begins with the *Questioner*, which is responsible for generating detailed questions based on various aspects of the image content. As shown in Fig. 6, the questions cover a wide range of topics, including objects, attributes, views, scenes, spatial relationships, text rendering, and world knowledge. For instance, questions like "What text is prominently displayed on the object?" and "What colours are used in the circular sign?".

Next, the *Answerer* utilizes VLMs such as InstructBLIP [13] or LLaVA [32] to respond to the generated questions based on the image content. For instance, some answers are the object is a "sun-like shape", made out of "paper or cardboard" and located "inside a building, likely a library or bookstore". These answers, however, are not always entirely accurate, reflecting the models' varying ability to understand the image. For instance, while both InstructBLIP and

Table 1: Quantitative comparison against SOTA methods on DOCCI dataset [33]. We report BLEU@4, ROUGE-L, METEOR, CIDEr, and average of caption length. The results from other methods are originally reported in [33].

Method	Eval mode	BLEU@4	ROUGE-L	METEOR	CIDEr	#Words
PaLI 5B [9]	Finetune on COCO	0.0	11.3	3.6	0.0	15.1
	Finetune on DOCCI	10.1	29.1	17.9	16.0	121.8
LLaVA-1.5 7B [32]	zero-shot	3.5	22.0	11.3	6.4	89.5
QAC-LLaVA (ours)		8.2	24.7	15.0	9.9	116.4
InstructBLIP (Vicuna-7B) [13]	zero-shot	3.5	20.5	10.6	5.9	84.4
QAC-InstructBLIP (ours)		7.6	24.6	13.5	9.4	115.8

LLaVA correctly identify the text "SUMMER READS", there are subtle differences in how they interpret elements like lighting and staircases. These discrepancies demonstrate the limitations of current VLMs in achieving perfect understanding.

Nevertheless, these answers serve their primary purpose, which is to guide the *Captioner* in generating detailed and accurate captions. The *Captioner* uses the information from the answers to produce captions that are both descriptive and informative. For example, the caption describes the scene as "a colorful sun-shaped sign reading 'SUMMER READS' in what appears to be a library or bookstore", incorporating specific details such as the object's material and the surrounding environment.

This case study provides strong evidence that our method effectively enhances the performance of pre-trained VLMs in a zero-shot setting. By using the *Questioner* and *Answerer* to extract relevant information and guide the *Captioner*, the approach enables the generation of more detailed and accurate image descriptions without requiring additional training data. Thus, it successfully improves the interpretive capabilities of VLMs.

4.3 Quantitative Comparison

Table 1 provides a comprehensive quantitative comparison of our method against other existing approaches. The results clearly demonstrate that QAC significantly enhances the performance of pre-trained models such as InstructBLIP [13] and LLaVA [32]. A particularly noteworthy observation is that our method improves all evaluation metrics, showing an overall boost in effectiveness. Additionally, our method allows the model to generate more detailed and longer image captions, with an average length of 115.8 and 116.4 words for each respective version. This substantial increase in caption length suggests that providing more detailed image information can greatly aid image-to-text generation tasks when leveraging pre-trained VLMs.

When we compare our method with other approaches that have been fine-tuned using datasets like COCO or DOCCI, the results become even more impressive. Our method outperforms PaLI [9], which has been fine-tuned on COCO,

**QAC-InstructBLIP**

1. The text is "SUMMER READS."
2. Yellow, orange, red, pink, and white.
3. Sun-like shape.
4. Paper or cardboard.
5. Inside a building, likely a library or bookstore.
6. "Join Us Book Club" and "new stuff we love."
7. Illustrations of an owl and a red animal.
8. Visible lighting.
9. Indications to join a book club.
10. Yes, there is a staircase present.

The image captures a vibrant, sun-shaped decoration prominently displaying the text "SUMMER READS" in colorful, hand-drawn letters. The sunburst features a mix of yellow, orange, red, pink, and white hues crafted from paper or cardboard, suggesting a cheerful, welcoming theme. This decoration hangs inside a library or bookstore, as indicated by signs promoting a book club and new reading materials displayed below. An inviting staircase with railings leads to an upper level, illuminated by overhead lighting, which adds to the cozy atmosphere. Additionally, charming illustrations of an owl and another red animal adorn the walls, further enhancing the inviting and creative vibe of the space.

QAC-LLaVA

1. "SUMMER READS"
2. Yellow, orange, red, pink
3. Sun
4. Paper or cardboard
5. Library or bookstore
6. "Join Us BOOK CLUB", "our stuff we love"
7. Owl illustration
8. Ceiling light fixture
9. No arrows, invitation to Book Club
10. Yes, in background

A colorful sun-shaped sign reading "SUMMER READS" is prominently displayed in what appears to be a library or bookstore. The sign, made of paper or cardboard, is a vibrant mix of yellow, orange, red, and pink, symbolizing the warmth of summer. Behind the sign, a staircase leads to an upper level, suggesting a cozy, multi-level space. On the walls, additional signs are visible, including one inviting people to join a "BOOK CLUB" and another saying "our stuff we love." A small owl illustration adds a whimsical touch to the scene. A ceiling light fixture shines softly above the staircase, contributing to the warm, inviting atmosphere. The display serves as a cheerful reminder of summer reading and community engagement within the space.

Fig. 6: A case study shows step-by-step of our approach. We show the input image and 10 questions generated by ChatGPT. Different versions of our approach including QAC-InstructBLIP and QAC-LLaVA, first answers those question and then gathers all answers to generate the detailed caption.

and closes the performance gap to PaLI fine-tuned on DOCCI to a greater extent than both InstructBLIP and LLaVA. While our method does not entirely surpass the performance of fine-tuned models, it is important to emphasize the considerable cost associated with fine-tuning. Finetuning requires extensive computational resources and high-quality datasets, which can be both expensive and time-consuming to acquire and process. In contrast, our method offers a more practical and efficient alternative, as it operates in a plug-and-play fashion, enabling it to be quickly integrated with any pre-trained VLMs without the need for extensive effort or resources.

4.4 Qualitative Comparison

Figure 7 presents examples of captions generated by different methods, offering a qualitative comparison of their performance. As shown, SmallCap [37] and EVCap [25] struggle to generate detailed captions, often failing to provide comprehensive descriptions of the image. In contrast, MiniGPT4 [60] manages



Ground-truth: A medium-long shot view of downtown Chicago, Illinois, with a view of the Chicago River. Along the near side of the river, a metal railway can be seen running horizontally. On the other side of the river, there are large buildings. Along the middle, there is a large rectangular building, and the same one is behind it. This building is made of light brown cement, and along this building there are squared windows. To the left of these buildings is one that is made up of glass, the bottom portion of this glass is covered in shade while the top portion of it is being lit up by the sun. In front of this building, there are trees that have orange leaves. To the right of the buildings in the middle, there is a skinny concrete building, and along the top portion of the building there are small cylindrical pillars that surround the building. To the right of this building are two glass-squared buildings that are made of glass and are reflecting the baby blue sky. Along the right side of the river there is a large square building that is submerged in shade, and in front of this building there is a bridge that runs horizontally, and is also covered in shade.

InstructBLIP: The image presents a tranquil cityscape, where nature and architecture coexist in harmony. The river, calm and reflective, serves as a mirror to the city's skyline, its surface undisturbed except for the occasional ripple. On the left side of the image, a large, ornate building stands tall. Its architecture is intricate, with a clock tower reaching towards the sky, punctuated by a pointed spire that adds to its grandeur. The building's facade is a mix of stone and glass, reflecting the sunlight and adding a touch of elegance to the scene. Opposite the ornate building, on the right side of the image, a modern skyscraper reaches for the heavens. Its glass exterior is a testament to contemporary architecture, reflecting the city's bustling life.

LLaVA: The image captures a serene urban scene, dominated by the presence of a river that cuts through the cityscape. The river, calm and reflective, serves as a mirror to the city's skyline, its surface undisturbed except for the occasional ripple. On the left side of the image, a large, ornate building stands tall. Its architecture is intricate, with a clock tower reaching towards the sky, punctuated by a pointed spire that adds to its grandeur. The building's facade is a mix of stone and glass, reflecting the sunlight and adding a touch of elegance to the scene. Opposite the ornate building, on the right side of the image, a modern skyscraper reaches for the heavens. Its glass exterior is a testament to contemporary architecture, reflecting the city's bustling life. The skyscraper's sleek lines contrast with the ornate details of the building across the river, creating a harmonious blend of old and new.

QAC-InstructBLIP: This image captures a picturesque view of downtown Chicago, showcasing a mix of historic and modern architecture along the Chicago River. The iconic Wrigley Building, with its distinguished clock tower, stands prominently, representing the city's rich architectural heritage. Adjacent skyscrapers with sleek glass facades rise against the blue sky, creating a striking contrast between old and new. A walkway runs parallel to the river, offering a scenic path for pedestrians, while the river itself reflects the towering structures, emphasizing the harmonious integration of urban design and natural elements. The late afternoon light casts a warm glow on the stone buildings, highlighting their intricate details and adding to the vibrant atmosphere of this bustling urban landscape.

QAC-LLaVA: This image showcases a prominent view of Chicago's iconic architecture along the Chicago River. In the foreground, the Wrigley Building stands tall with its historic clock tower, a symbol of the city's architectural heritage since 1920. Flanking it are modern glass skyscrapers, reflecting Chicago's balance between its storied past and cutting-edge present. The Chicago Riverwalk is visible along the right, a popular public space offering stunning views of the city's skyline. The Tribune Tower, another landmark with gothic architecture, adds to the blend of old and new. The calm waterway mirrors the buildings, highlighting Chicago's reputation as a global center for architecture and design. The scene captures the city's dynamic urban landscape, blending history, innovation, and the beauty of its riverside location.

MiniGPT4: The image shows a view of a river running through the city, with tall buildings on either side. The buildings are made of concrete and steel, with large windows and pointed roofs. The river is clear and blue, with boats and other watercraft moving along it. The sky is a bright blue, with a few clouds in the distance. The sun is shining, casting a warm glow over the scene. The overall atmosphere is peaceful and serene.

SmallCap: A large body of water with tall buildings in the background.

EVCap: A large body of water next to tall buildings.



Ground-truth: A view of a black Dodge Ram truck parked on pavement with two black cars lined up behind it. There are diagonal white lines on the pavement to the right of them. The cars are facing the front, and all have reflections on their windshields. There is a stoplight overhead on the left. There is a strip of buildings on the left, alongside the road. The first one has flowers around the top of the striped awning. The tall building in the middle with red brick has ladders on the front of it and graffiti sprayed on the side of it. There is part of a tree visible on the right, next to the third parked car. There are more tall buildings visible in the background.

InstructBLIP: The image shows a city street scene with a row of parked vehicles. Behind these vehicles, there are more cars parked, and the street is lined with buildings that have storefronts and awnings. The architecture suggests an urban environment, and there are traffic lights visible, indicating a regulated traffic area. The sky is partly cloudy, suggesting it might be a cool or overcast day.

LLaVA: The image depicts a city street scene with a row of parked vehicles. The most prominent vehicle is a large, black pickup truck with a distinctive front grille design, which suggests it might be a luxury or high-performance model. The truck is parked alongside other vehicles, including a smaller car and another truck. The street is lined with buildings, and there are traffic lights visible, indicating an urban setting. The presence of a fire escape on one of the buildings and the architecture of the buildings suggest this could be a city in the United States.

QAC-InstructBLIP: This image captures a street scene in an urban area, possibly in New York City, recognizable by its characteristic architecture and graffiti-covered walls. In the foreground, there's a lineup of black vehicles, led by a customized Ram 1500 truck. The truck features an aggressive, armored front bumper, designed for off-road use and heavy-duty applications, giving it a formidable presence. Following the truck are other black SUVs, adding to the tough, unified look. The background reveals classic New York City buildings with red brick facades, fire escapes, and storefronts, while overhead traffic lights and street signs emphasize the city's dynamic, fast-paced environment.

QAC-LLaVA: The image captures an urban street scene with three black vehicles parked in a row along the curb. These vehicles include a Ram truck at the front and two SUVs trailing behind. The street is lined with a variety of buildings, indicating a mixed-use neighborhood possibly in a city. Some buildings have notable fire escapes on their facades, while others display commercial signage. Visible graffiti adds to the urban aesthetic. A traffic signal hangs above the street, and there are trees providing some greenery in the scene. Sidewalks are bordered with flowers and decorative elements, enhancing the urban environment. The overall atmosphere is a blend of modern urban life with hints of historical architecture.

MiniGPT4: The image depicts a row of three dark-colored vehicles parked in a line on a city street. The vehicles are parked along the curb in a clearly marked parking area. The parked vehicles seem to be part of a bustling, well-visited area.

SmallCap: A black truck is driving down the street.

EVCap: A row of black trucks parked on a city street.

Fig. 7: Generated captions by compared methods. Captions generated by QAC-InstructBLIP and QAC-LLaVA (both are ours) are more detailed than those by other methods. They also contain world knowledge (highlighted in red) which is closed to the ground-truth captions.

to capture some level of detail, but its performance is still limited compared to the more sophisticated methods. InstructBLIP, LLaVA, QAC-InstructBLIP, and QAC-LLaVA perform noticeably better, generating captions that are richer in detail.

Although it may be challenging to definitively determine which method among InstructBLIP, LLaVA, QAC-InstructBLIP, and QAC-LLaVA performs best, especially since VLMs are generally capable of capturing basic image information, our methods show a clear advantage. Specifically, QAC-InstructBLIP and QAC-LLaVA enrich the captions with external knowledge, bringing them closer to the ground-truth descriptions than the other methods as highlighted in red. This ability to incorporate world knowledge demonstrates that our question-answer-caption (QAC) framework significantly enhances caption quality in a meaningful way. The benefits of our approach are further supported by the quantitative results discussed earlier, confirming that this simple yet effective method boosts the level of detail in image captions, particularly in zero-shot scenarios.

4.5 Plug-and-Play Ability

Our method can be implemented on pre-trained VLMs in a zero-shot manner, enabling it to function as a plug-and-play module compatible with any pre-trained VLMs. To verify this capability, we implement it on InstructBLIP and LLaVA. As demonstrated in Figures 6 and 7, as well as Table 1, our method consistently boosts the performance of both InstructBLIP and LLaVA.

However, we observe variations in performance between QAC-InstructBLIP and QAC-LLaVA models, which could be attributed to differences in the original capabilities of each pre-trained VLM and the prompts used. Firstly, as shown in Figure 6, the two models respond differently to the same questions, indicating variations in their ability to comprehend and interpret image content. For instance, InstructBLIP might identify the owl illustration more effectively, while LLaVA might provide a more detailed description of the staircase. Secondly, our method employed a simple prompt to interact with the VLMs, which might not be robust enough to fully guide the models. This remains a room where further refinement could lead to improved performance, suggesting that prompt engineering could enhance the overall effectiveness of our approach. The potential impact of more sophisticated prompts remains an avenue for future research. Additionally, exploring our method’s compatibility with other pre-trained VLMs is crucial to further elaborate on its efficacy. Despite these considerations, current observations suggest that our method is beneficial in enhancing both world knowledge and generating more detailed captions.

4.6 Impact of Questioner

To understand the impact of different *Questioners* on our approach, we conduct an ablation study by comparing the effectiveness of using InstructBLIP, LLaVA, and ChatGPT as the *Questioner* module (see Table 2). When using

Table 2: Ablation study on the effect of different generators.

Method	Questioner	BLEU@4	ROUGE-L	METEOR	CIDEr	#Words
QAC-InstructBLIP	ChatGPT	7.6	24.6	13.5	9.4	115.8
QAC-LLaVA		8.2	24.7	15.0	9.9	116.4
QAC-InstructBLIP	InstructBLIP [13]	5.2	21.1	11.7	6.7	92.8
QAC-LLaVA		5.4	22.3	12.0	6.9	91.3
QAC-InstructBLIP	LLaVA [32]	5.6	22.4	11.9	6.4	93.7
QAC-LLaVA		6.2	22.7	12.5	7.6	95.1

InstructBLIP and LLaVA, the generated questions are simpler and lack of insights, resulting in less detailed answers and ultimately reducing the quality of the captions. In contrast, using ChatGPT as the *Questioner* leads to more contextually rich and comprehensive questions, which will provide the *Captioneer* with more information to create detailed and accurate captions. Our ablation study shows that the choice of *Questioner* significantly affects the performance of our proposed method.

We also conduct an ablation study to examine the impact of the number of questions generated by the *Questioner* on the overall performance of our method. In this experiment, we use ChatGPT as the *Questioner* and generate a number of 1, 5, 10, 15, and 20 questions, respectively. The BLEU@4 and CIDEr scores are plotted in Figure 8. We see that when a smaller number of questions is used (i.e., 1 or 5), the answers are less comprehensive, leading to captions that missed certain details and nuances of the image content. On the other hand, as the number of questions increases (i.e., 10), the *Questioner* is able to extract more detailed information, enabling the *Captioneer* to produce more informative and accurate captions. However, we notice that beyond a certain threshold (i.e., 15, 20), the additional questions provide diminishing returns, as the extra information becomes repetitive and does not significantly improve the final captions. Therefore, finding an optimal number of questions is crucial for balancing detail and efficiency, ensuring the most effective enhancement of VLM performance.

5 Discussion

Our method, QAC, delivers several advantages. (1) **Adaptability:** Due to the continuing advancement of both vision and language models, our proposed approach can be adopted by any state-of-the-art VLMs effortlessly in order to enhance the overall performance. (2) **Interpretability:** The generated questions set are easy to understand and portray multiple perspectives of the visual content. With our own knowledge, human can comfortably verify the relevance of the questions related to the input photo as well as how accurate of the corresponding answers. (3) **Robustness:** By leveraging a wide range of descriptive

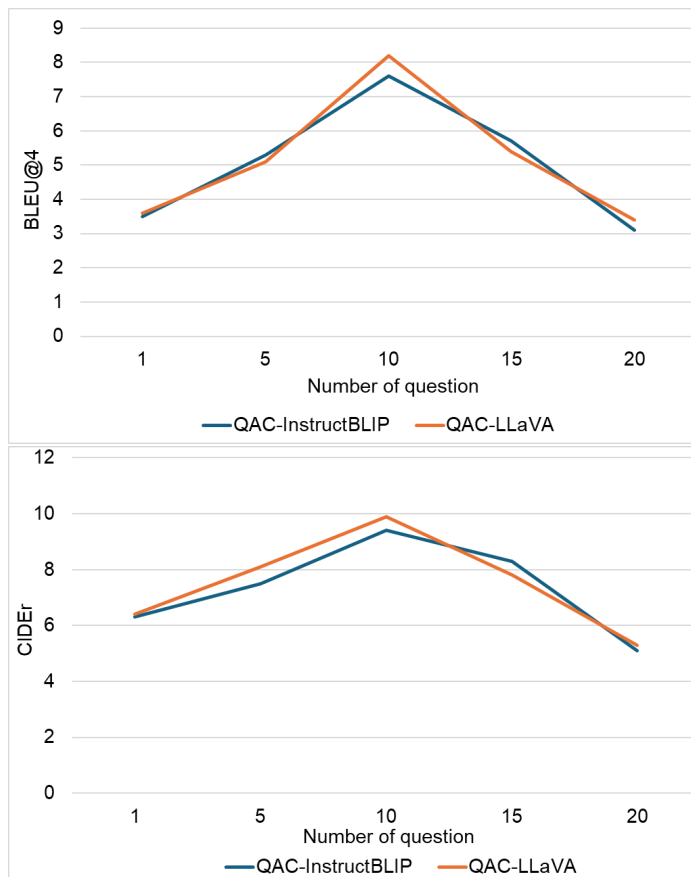


Fig. 8: Ablation study on the impact of the number of question. When fewer questions or more questions are asked, the performance degrades in both BLEU@4 (above) and CIDEr (below) scores.

information, our approach produces accurate and detailed descriptions that are comparable to human annotations.

Nonetheless, we recognize some failure cases that reveal limitations of our method. Firstly, our method heavily relies on the performance of the *Questioner*. As highlighted in Table 2 above, low-quality *Questioner* leads to diminished performance. Specifically, when the *Questioner* generates questions that are either irrelevant or overly specific, especially in complex images with intricate details, the answers do not contribute meaningfully to the overall understanding. This results in captions that are cluttered with unnecessary information, reducing their clarity and coherence.

Secondly, the performance of the *Answerer* significantly influences the quality of the final captions. When the *Answerer* struggles to provide accurate responses, the generated captions are often incomplete or misleading. In fact, when

we used ChatGPT as the *Answerer* on samples with low quantitative scores from QAC-InstructBLIP and QAC-LLaVA, the performance improved noticeably, indicating the importance of the interpretive ability of VLMs. However, since our goal is to develop a plug-and-play method adaptable to any pre-trained VLMs, we do not delve deeply into this issue.

Thirdly, our method does not intentionally control the hallucination tendencies inherent in VLMs. As a result, the generated captions sometimes contain irrelevant information that does not correspond to the image. Addressing this limitation would require improving the fairness and accuracy of LLMs, which falls beyond the scope of our current work.

6 Conclusion

In this paper, we present a novel method that leverages a *Question – Answer – Caption* framework to enhance the performance of pre-trained VLMs in zero-shot detailed image captioning task. Our approach effectively utilizes the *Questioner* to generate questions to extract detailed information from the image. Then, the answers by the *Answerer* guide the *Captioner* to produce rich and informative captions. Through extensive experiments with InstructBLIP and LLaVA, we demonstrate that our method improves their captioning capabilities. Moreover, our method serves as a plug-and-play module that is compatible with any pre-trained VLMs. While we identified certain limitations, they highlight room for future refinement, providing a pathway for further advancements in image captioning.

Acknowledgements: This work was supported by JSPS/MEXT KAKENHI Grant Numbers JP24K20830, and ROIS NII Open Collaborative Research 2024-24S1201.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **35**, 23716–23736 (2022)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6077–6086 (2018)
4. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)

5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: 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. pp. 3558–3568 (2021)
6. Chen, D., Cahyawijaya, S., Ishii, E., Chan, H.S., Bang, Y., Fung, P.: The pyramid of captions. arXiv preprint arXiv:2405.00485 (2024)
7. Chen, G.H., Chen, S., Zhang, R., Chen, J., Wu, X., Zhang, Z., Chen, Z., Li, J., Wan, X., Wang, B.: Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. arXiv preprint arXiv:2402.11684 (2024)
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
9. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
10. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) **2**(3), 6 (2023)
12. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **24**(240), 1–113 (2023)
13. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
15. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:2111.11431 (2021)
16. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. *Advances in Neural Information Processing Systems* **36** (2024)
17. Garg, R., Burns, A., Ayan, B.K., Bitton, Y., Montgomery, C., Onoe, Y., Bunner, A., Krishna, R., Baldrige, J., Soricut, R.: Imageinwords: Unlocking hyper-detailed image descriptions. arXiv preprint arXiv:2405.02793 (2024)
18. Gaur, M., Tapaswi, M., et al.: No detail left behind: Revisiting self-retrieval for fine-grained image captioning. arXiv preprint arXiv:2409.03025 (2024)
19. Ge, Y., Zeng, X., Huffman, J.S., Lin, T.Y., Liu, M.Y., Cui, Y.: Visual fact checker: Enabling high-fidelity detailed caption generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14033–14042 (2024)
20. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14281–14290 (2024)

21. Hsieh, Y.G., Hsieh, C.Y., Yeh, S.Y., Béthune, L., Ansari, H.P., Vasu, P.K.A., Li, C.L., Krishna, R., Tuzel, O., Cuturi, M.: Graph-based captioning: Enhancing visual descriptions by interconnecting region captions. arXiv preprint arXiv:2407.06723 (2024)
22. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4565–4574 (2016)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
24. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
25. Li, J., Vo, D.M., Sugimoto, A., Nakayama, H.: Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13733–13742 (2024)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
28. Li, X., Zhang, F., Diao, H., Wang, Y., Wang, X., Duan, L.Y.: Densefusion-1m: Merging vision experts for comprehensive multimodal perception. arXiv preprint arXiv:2407.08303 (2024)
29. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26763–26773 (2024)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36** (2024)
33. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., et al.: Docci: Descriptions of connected and contrasting images. arXiv preprint arXiv:2404.19753 (2024)
34. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 647–664. Springer (2020)
35. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6967–6977 (2023)

36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: Smallcap: lightweight image captioning prompted with retrieval augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2840–2849 (2023)
38. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
40. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **35**, 25278–25294 (2022)
41. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
42. Shi, B., Wu, Z., Mao, M., Wang, X., Darrell, T.: When do we not need larger vision models? arXiv preprint arXiv:2403.13043 (2024)
43. Singla, V., Yue, K., Paul, S., Shirkavand, R., Jayawardhana, M., Ganjdanesh, A., Huang, H., Bhatele, A., Somepalli, G., Goldstein, T.: From pixels to prose: A large dataset of dense image captions. arXiv preprint arXiv:2406.10328 (2024)
44. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14398–14409 (2024)
45. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
46. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: The Twelfth International Conference on Learning Representations (2023)
47. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
48. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. *Communications of the ACM* **59**(2), 64–73 (2016)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26700–26709 (2024)

51. Wang, W., Ren, Y., Luo, H., Li, T., Yan, C., Chen, Z., Wang, W., Li, Q., Lu, L., Zhu, X., et al.: The all-seeing project v2: Towards general relation comprehension of the open world. arXiv preprint arXiv:2402.19474 (2024)
52. Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., Chen, Z., Li, H., Zhu, X., Cao, Z., et al.: The all-seeing project: Towards panoptic visual recognition and understanding of the open world. arXiv preprint arXiv:2308.01907 (2023)
53. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706 (2019)
54. Xu, R., Yao, Y., Guo, Z., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Sun, M., Huang, G.: Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. arXiv preprint arXiv:2403.11703 (2024)
55. You, H., Guo, M., Wang, Z., Chang, K.W., Baldrige, J., Yu, J.: Co-bit: A contrastive bi-directional image-text generation model. arXiv preprint arXiv:2303.13455 (2023)
56. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* **2**, 67–78 (2014)
57. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
58. Yu, Q., Sun, Q., Zhang, X., Cui, Y., Zhang, F., Cao, Y., Wang, X., Liu, J.: Capsfusion: Rethinking image-text data at scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14022–14032 (2024)
59. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6720–6731 (2019)
60. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)