

VEGCAP: A Visual Entity-Guided Image Caption Generation method for Visual Question Answer

ChuanZhi Zhang¹, MingJie Wang^{1,2}, JuXiang Zhou^{1,2†}, ZeGuang Lu³, JiaTian Mei^{1,2}, YaJie Gu¹

¹Key Laboratory of Education Informatization for Nationalities, Ministry of Education, Yunnan Normal University, Kunming, 650500, China

²Yunnan Key Laboratory of Smart Education, Yunnan Normal University, Kunming, 650500, China

³National Academy of Guo Ding Institute of Data Sciences, Beijing, 100080, China

Abstract

The task of Visual Question Answering (VQA), particularly in datasets like OK-VQA, demands the integration of external knowledge to provide accurate answers. In recent years, approaches that leverage large language models as implicit knowledge bases have demonstrated remarkable progress. However, a significant challenge for these approaches is effectively bridging the gap between visual data and language models. Representing image content solely through image captions and object labels may overlook crucial information and introduce noise, failing to adequately capture the complexity and details of the image, thereby affecting the accuracy of the model's responses. To address this issue, this paper proposes VEGCAP, a visual entity-guided image captioning method. By analyzing natural language questions, VEGCAP accurately identifies the visual entities in the image that need to be emphasized, generating more targeted image descriptions. Compared to general image captioning methods, VEGCAP effectively minimizes interference from irrelevant information and highlights key details related to the question. Experimental results demonstrate that the proposed method achieves state-of-the-art accuracy on knowledge-based VQA datasets, OK-VQA and A-OKVQA, reaching 60.9% and 59.2% on the validation sets, respectively.

Keywords: Visual Question Answering; Large Language Models; Image Captioning; Few-Shot Learning; Visual Entity-Guided

1. Introduction

The task of Visual Question Answering (VQA) was first introduced by Antol et al [1]. In this task, a model receives a natural language question Q and an image I as inputs, and generates an answer A based on the information contained in both. VQA tasks can be categorized into two types: one that utilizes information from the image to perform fine-grained recognition, scene recognition, attribute classification, and counting; and another that requires reasoning

[†]Corresponding author: JuXiang Zhou (Email: zhoujuxiang@ynnu.edu.cn; ORCID:0000-0003-2693-2204)

based on knowledge beyond the image to generate an answer. Thanks to advances in deep learning, state-of-the-art methods have achieved satisfactory results in the first type of task. However, knowledge-based Visual Question Answering [2] sets higher demands for VQA tasks, as it involves external knowledge for reasoning.

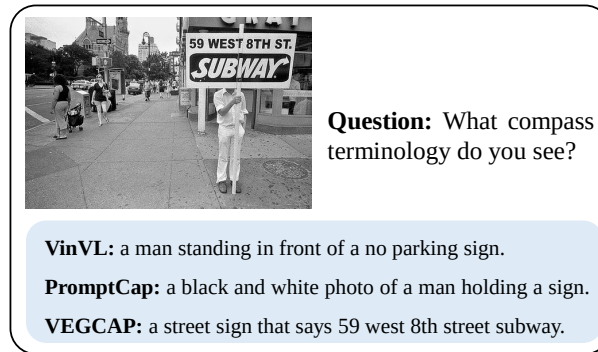


Figure 1: A test sample from the OK-VQA dataset, which includes an image, a question about the image, and the image’s caption (blue), allowing the large language model to utilize this textual information to answer the posed question.

Some existing methods retrieve external knowledge from a variety of sources, including knowledge graphs [3]-[5], Wikipedia [4]-[6], and Google Search [7]-[9]. However, retrieving the necessary knowledge often introduces a significant amount of irrelevant information, which can negatively impact the model’s training and reasoning. To address this, in addition to using explicit knowledge bases, some studies leverage pre-trained Large Language Models (LLMs) as implicit knowledge bases to perform knowledge retrieval and reasoning, enabling the model to answer questions more effectively. For example, PICa [10] utilizes the VinVL model [25] to convert images into image captions, then feeds the test sample’s question and image caption, along with a few training samples containing questions, captions, and answers, into GPT-3 [15]. GPT-3 [15] can adapt to new tasks based on the training samples and leverage its powerful reasoning capabilities to answer questions. Although a few context examples can effectively stimulate the performance of GPT-3 [15], the captions generated by the VinVL model are insufficient to cover the key information needed to answer the questions. PromptCap [11] uses the question as a prompt to guide the model’s attention toward relevant entities, but the generated captions for these entities lack sufficient detail. Prophet [12] mitigates this issue by introducing candidate answers for the question through a simple VQA model, which improves performance. Although the enhancements to the MCAN model [33] in the Prophet [12] paper are significant, relying solely on images and questions during training results in a limited acquisition and understanding of external knowledge. Additionally, the VinVL model can only describe the main subjects of an image. When questions involve background objects or text on objects, GPT-3 [15] struggles to provide accurate answers. For example, as shown in Figure 1, when asked, ‘What compass terminology do you see?’, the caption generated by VinVL [25] is ‘a man standing in front of a no parking sign.’ which is clearly incorrect. The caption from PromptCap [11] only mentions ‘a man holding a sign,’ which also fails to provide the critical information needed for the large language model to answer the question. Furthermore, the mention of ‘no parking sign’ can mislead the model, causing it to make erroneous judgments when answering. In contrast, the method proposed in this paper provides a detailed description of the text information on the sign, offering

the model clearer context. This level of detail is essential for preventing faulty reasoning.

To address the issues mentioned above, this paper introduces a visual entity-guided image captioning method called VEGCAP. This approach accurately identifies the visual entities in an image that require detailed description based on the content of the question, enabling the extraction of more critical visual information. This ensures that the large language model can better understand the image content and provide more accurate answers. Additionally, the improved CapMCAN model integrates these captions as auxiliary information during training, enriching the visual representation of the image while leveraging the implicit external knowledge embedded in the captions. This method improves the model's ability to handle complex visual question answering tasks, particularly those that involve external knowledge.

The contributions of this paper are as follows:

1. This paper introduces a visual entity-guided image captioning method (VEGCAP), which analyzes the content of the question to accurately identify relevant entities in the image. By generating captions closely aligned with the question, VEGCAP improves the connection between visual and textual information.
2. We enhance the original MCAN model and introduce the CapMCAN model. This model integrates VEGCAP captions as auxiliary information and employs a co-attention mechanism to fuse multimodal image features, resulting in more accurate candidate answers and context examples.
3. From the perspective of operational cost, this paper uses the much smaller Mistral 7B [16] model, which has significantly fewer parameters than GPT-3's 175B, as the question-answering model. Experiments on the OK-VQA [2] and A-OKVQA [19] datasets demonstrate that the open-source Mistral model can be effectively applied to knowledge-based visual question answering tasks without sacrificing too much accuracy. Furthermore, the proposed method allows the Mistral model to achieve performance comparable to that of Prophet.

2. Related Work

Knowledge-based Visual Question Answering: The key to knowledge-based visual question answering lies in the effective acquisition and integration of knowledge. Previous works have primarily sourced the necessary knowledge from various external knowledge resources, such as ConceptNet [20], Wikipedia [21], and Google Search [7], and fed this acquired knowledge along with images and questions into the model to generate answers. With the advancement of large language models, which have excelled in various natural language tasks, some methods have leveraged their powerful knowledge reasoning capabilities to directly predict answers [10][11][12]. Other studies [22][23] have utilized GPT-3 [15] as an implicit knowledge base to retrieve implicit knowledge, enhancing the accuracy of subsequent answers. Building on the work of [12], this paper enhances the image captions used to represent visual content, further unlocking the potential of large language models.

Visual-Language Model: Through training on large-scale image-text pairs, visual-language models [24]-[31] learn associations between images and text, demonstrating strong transfer capabilities and achieving state-of-the-art performance after fine-tuning on various downstream tasks. Models like Flamingo [29] and BLIP-2 [30] incorporate large language models as core components, aligning image and text modalities by freezing the language model while tuning the visual encoder. OFA [31] adopts a Transformer [32] architecture to integrate image and text

information within a unified encoder-decoder format, eliminating the need for different model layers for various tasks.

Large Language Model: Large language models [15]-[18] are typically Transformer-based [32] architectures with billions or even hundreds of billions of parameters. Trained on massive amounts of text data, these models exhibit strong capabilities in natural language understanding and generation. Their versatility makes them well-suited for a wide range of language-related tasks, including question answering, text generation, text summarization, and sentiment analysis. Due to GPT-3's [15] powerful knowledge retrieval and reasoning abilities, PICa [10] and its subsequent works [11]-[14] enable GPT-3 [15] to tackle visual question answering tasks by converting the visual information from images into textual information. Bridging the gap between visual and natural language and better connecting image and language models are key factors in enhancing the performance of large language models.

3. Method

The method proposed in this paper is based on the framework of Prophet [12] and aims to address knowledge-based visual question answering tasks by leveraging large language models for few-shot context learning. As illustrated in Figure 2, the method is divided into two stages: In Stage 1, the inputs for the large language model are prepared. These inputs include the question-guided image captions generated by the PromptCap model, the visual entity-guided image captions produced by VEGCAP, as well as the candidate answers and context examples obtained from the CapMCAN model. In Stage 2, the training examples most similar to the test example are selected as learning samples and provided to the Mistral large language model alongside the test example. Using the input information, combined with its own knowledge, the model generates the most accurate answer.

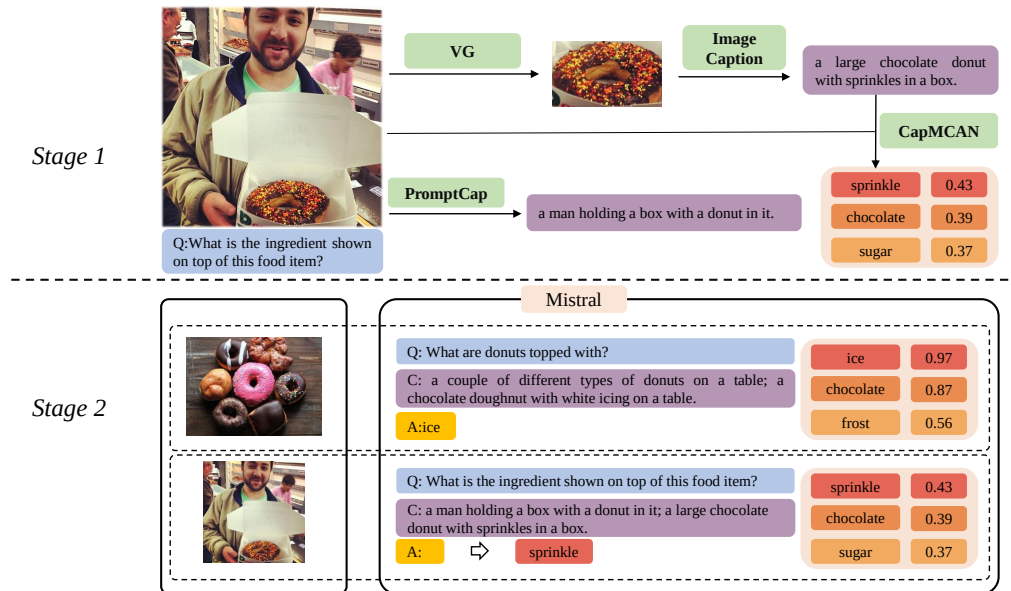


Figure 2: Our model architecture consists of two main steps: data preparation and model inference.

3.1. Image Caption

When using large language models for visual question answering tasks, accurate image captions are essential, as they contain all the key visual information relevant to the question. Any omission of important visual details can significantly impact the quality of the model's generated answers. To ensure the model's effectiveness within limited computational resources and time costs, this paper selects different variants of the OFA model [31] to separately handle the tasks of visual grounding and image caption. This strategy allows the models to leverage their strengths in specific tasks, thereby improving the overall efficiency and accuracy of the tasks.

To validate the effectiveness of the proposed visual entity-guided image caption method, we employed two different strategies to generate image captions. First, we generated a 'general caption' for each image without considering specific questions. This caption primarily summarizes the main visual content in the image. However, the content covered by these general captions may not directly be applicable to answering questions related to the image. Therefore, we further introduced visual grounding techniques to identify specific objects or characters mentioned in the question, generating more precise captions that focus on these targeted elements.

To ensure that our research findings are not influenced by different image caption models, we conducted a comparative analysis between the method proposed in this paper and the general captions generated by OFA [31] as well as the captions generated by VinVL [25].

3.2. Candidate Answers and Context Examples

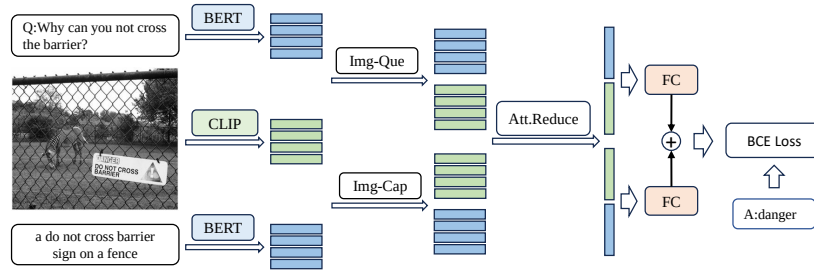


Figure 3: CapMCAN model architecture.

In this section, we discuss the generation of candidate answers and context examples, which are essential for the visual question answering process. The CapMCAN model in Figure 2 is used to generate candidate answers and select relevant context examples. To improve the accuracy of generating candidate answers, we utilize the previously mentioned image descriptions based on visual grounding to enhance the visual representation of the image, as shown in the model structure in Figure 3. Since pre-trained vision-language models have been shown to incorporate some world knowledge, the introduction of image descriptions not only enriches the visual representation of the image but also brings in implicit external knowledge, which helps to enhance the overall performance of the model.

In the original MCAN model [33](see Figure 4a), question features extracted by BERT are fused with image visual features extracted by CLIP using guided attention (GA), as shown in Figure 5a. Building on this, we propose the Img-Cap module in Figure 4b, forming a dual-branch CapMCAN model. In this model, after passing through a self-attention layer, the image description features are fused with the image features using co-attention (CA), as illustrated in

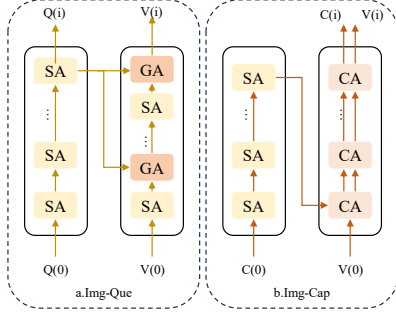


Figure 4: Img-Que(a) and Img-Cap(b).

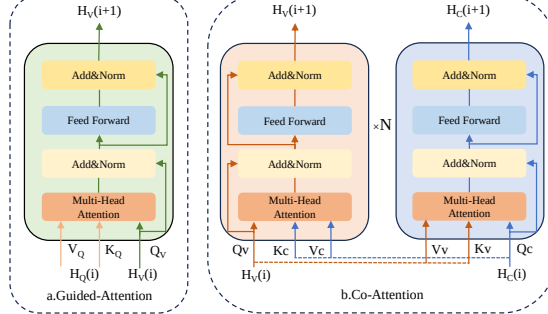


Figure 5: GA(Guided-Attention) and CA(Co-Attention).

Figure 5b. To effectively merge the semantically rich features, the image description features are first reduced in dimensionality through an attention reduction module, as described by the following equations:

$$\alpha = \text{softmax}(MLP(C)) \quad (1)$$

$$\tilde{c} = \sum_{i=1}^m \alpha_i c_i \quad (2)$$

The image description C is processed through two fully connected layers to compute an attention score α , which is then used to weighted aggregate the image description features into a one-dimensional vector $\tilde{c}$ of length d . The fused question features, image features, and image description features are then reduced in dimensionality and summed before being passed through a binary cross-entropy loss function with the answer vector.

For a trained model, a test example (v_i, q_i, c_i) is input into the CapMCAN model, where v_i, q_i, c_i represent the image, the question, and the caption, respectively. The model passes through the classification head and outputs the final prediction vector $y \in R^{1 \times L}$, where L is the length of the answer set. The top K highest-scoring answers are selected as candidate answers, as given by the following formula:

$$I_{AC} = \arg \text{TopK}_{y[j]} \quad (3)$$

Let I_{AC} be the index set of the top K candidate answers. The final candidate answer set D includes the candidate answers w_j and their corresponding confidence scores $y[j]$:

$$D = \{(w_j, y[j]) \mid j \in I_{AC}\} \quad (4)$$

Previous studies have shown that good context examples can significantly enhance the few-shot ability of large language models. In this paper, we use the fused features before the classification layer of the CapMCAN model to select context examples that may have similar answers. By calculating the cosine similarity between the test sample z and training samples z_i , we select the most similar context examples, as shown in the following formula:

$$I_{AE} = \arg \text{TopN}_{i \in \{1, 2, \dots, M\}} \frac{z^T z_i}{\|z\|_2 \|z_i\|_2} \quad (5)$$

Here, I_{AE} is the index set of training samples that are similar to the test sample z .

3.3. Question Answering

As shown in Figure 2, in the second step, we utilize the Mistral 7B model as our question-answering model. Mistral 7B is a large language model with 7 billion parameters, outperforming Llama2-13B [17] across all test sets. For the input to the question-answering model, we generate candidate answers and context examples using the CapMCAN model. For image descriptions, we adopt the general descriptions provided by the PromptCap method, which focus on the entities relevant to the question, and complement these with the image location-based descriptions proposed in this paper, which offer more detailed information about the entities in question. The training examples include both the corresponding image descriptions and questions; however, unlike the training samples, the answer sections of the test samples are left blank. The training examples include the following information:

Caption: c_i
Question: q_i
Candidates: $w_{i_1}(s_{i_1}), w_{i_2}(s_{i_2}), \dots, w_{i_k}(s_{i_k})$
Answer: a_i

4. Experiments

We evaluated the performance of the proposed method on two popular knowledge-based visual question answering datasets: OK-VQA and A-OKVQA. Through ablation experiments on the OK-VQA dataset, we explored the effectiveness of the proposed approach. The results indicate that our method performed exceptionally well on both datasets, validating its effectiveness in knowledge-based visual question answering tasks.

4.1. Datasets

OK-VQA [2], as a commonly used knowledge-based dataset, includes a training set of 9,000 image-question pairs and a test set of 5,000 image-question pairs. To ensure that external knowledge is required to answer the questions, all questions were manually filtered, and each image in the dataset has one question and ten human-annotated ground truth answers corresponding to it.

A-OKVQA [19], as a follow-up to OK-VQA, is currently the largest knowledge-based VQA dataset. This dataset includes a training set of 17,000 examples, a validation set of 1,000, and a test set of 7,000. Similar to OK-VQA, direct answer evaluation requires each question to have ten open-ended answers, while multiple-choice evaluation requires the model to select the correct answer from four options. Compared to OK-VQA, A-OKVQA not only tests the model’s ability to integrate knowledge but also assesses the model’s capability to handle ambiguous or polysemous questions.

4.2. Evaluation

We used the soft accuracy proposed in VQAv2 [1] to evaluate the model’s performance on direct answers for OK-VQA and A-OKVQA:

$$Acc(ans) = \min\left(1, \frac{\{\# \text{ humans provided ans}\}}{3}\right) \quad (6)$$

If at least three of the ground truth answers match the predicted answer, the prediction is considered 100% accurate. For the multiple-choice task in A-OKVQA, this paper follows the strategy outlined in [19], projecting the predicted answer onto the nearest available option.

4.3. Implementation Details

In this paper, we utilized the visually-grounded variant of OFA-large to identify the visual entities in the image that require detailed descriptions, and we employed the image caption variant of OFA-huge to generate detailed descriptions of these entities. Prophet [12] used the MCAN-large [33] model as the VQA model to generate candidate answers and answer-aware examples. To better leverage the implicit external knowledge contained in the image captions, we proposed the CapMCAN model, which integrates a branch for merging images and image captions. The CapMCAN model utilizes the CLIP RN50×64 visual encoder to process image information, while the questions and captions are encoded using BERT-large. In the fine-tuning phase, the MCAN-large model weights trained on the VQAv2 dataset in Prophet [12] are used to initialize the CapMCAN model, which is then fine-tuned on the OK-VQA dataset. In the experiment, the batch size is set to 64. Initially, the Adam optimizer is used with a learning rate of 0.001 to update only the newly added layers, with a training period of 2 epochs. Subsequently, the AdamW optimizer is employed to fine-tune the entire model, with a base learning rate set to 7×10^{-5} , which decays by a factor of 0.2 at the 5th epoch, and a maximum training period of 6 epochs.

In the prompting phase, the number of candidate answers K is set to 10, the number of context examples N is set to 20, and the number of queries T is set to 5 as default settings. The version of Mistral used in the experiments is Mistral-7B-v0.1, with a sampling temperature set to 0. All experiments, including the training of the CapMCAN model and the inference of the Mistral model, were conducted on Nvidia A100 GPUs.

4.4. Main Results

Table 1 compares existing methods with the proposed approach on the OK-VQA test set, detailing the knowledge sources for each method. Retrieving the necessary knowledge for answering questions from vast knowledge bases typically requires extensive training and may introduce irrelevant information. As shown in the table, some current methods tend to treat large language models as implicit knowledge bases and question-answering systems. However, how to enable large language models to fully understand image content remains an unresolved issue.

PICa [10] uses image captions and object labels to represent images, but sometimes the entities described in the image captions are unrelated to the questions posed. To address this issue, PromptCap [11] controls the visual content to be described by utilizing the questions during image caption generation. Prophet [12] enhances GPT-3’s input by incorporating candidate answers and selecting answer-aware examples, effectively tapping into the potential of large language models. However, for large language models, the textual representation of images is crucial. Most of the aforementioned methods use general image captions that cover the entire image, which may overlook critical details.

Table 1: The comparison of results from existing models on the OK-VQA dataset.

Method	Knowledge Resources	Accuracy(%)
ConceptBERT[3]	ConceptNet	33.7
KRISP[4]	Wikipedia+ConceptNet	38.4
MAVEx[5]	Wikipedia+ConceptNet+Google Images	39.4
PICa-Base[10]	Frozen GPT-3 (175B)	43.3
PICa-Full[10]	Frozen GPT-3 (175B)	48.0
KAT(Single)[22]	Wikidata+Frozen GPT-3 (175B)	53.1
KAT(Ensemble)[22]	Wikidata+Frozen GPT-3 (175B)	54.4
REVIVE(Single)[23]	Wikidata+Frozen GPT-3 (175B)	56.6
REVIVE(Ensemble)[23]	Wikidata+Frozen GPT-3 (175B)	58.0
Flamingo-80B[29]	Flamingo (80B)	57.8
REAVL[39]	image 4M + Wikidata	57.7
CVLM[[34]	Wikidata	58.9
PromptCap[11]	Frozen GPT-3(175B)	60.4
TOA[35]	ChatGPT	60.6
MM-Reasoner[40]	GPT-4	60.8
Prophet[12]	Frozen GPT-3(175B)	61.1
ours	Mistral(7B)	<u>60.9</u>

The proposed method aims to address the shortcomings of general image captions by specifically describing the image entities relevant to the questions. This targeted approach not only generates more accurate candidate answers but also provides more relevant context examples, thereby maximizing the few-shot learning capabilities of large language models and improving overall performance on knowledge-based visual question-answering tasks. Experimental results demonstrate that our proposed method achieves performance comparable to that of GPT-3, which has 175 billion parameters, while utilizing an open-source model with only 7 billion parameters.

Unlike the OK-VQA dataset, A-OKVQA requires answers to questions to rely more heavily on external knowledge and complex reasoning based on visual cues. Table 2 presents a comparison of various methods, including the proposed approach, on the A-OKVQA dataset. The results indicate that our proposed method outperforms most existing approaches and achieves higher accuracy compared to the Prophet method. Experimental findings demonstrate that our approach effectively bridges the gap between images and language models, further enhancing the capabilities of large language models.

4.5. Ablation Experiments

We conducted ablation experiments on the OK-VQA dataset to assess the effectiveness of the proposed method. By systematically removing key components of the model, the experiments reveal the impact of each module on the final performance. Specifically, the experiments analyze the contributions of the visual entity-guided image captions (VEGCAP), PromptCap captions, and the candidate answers and context examples generated by the CapMCAN model by removing them one by one. The results indicate that removing any single module leads to a decline in model performance. This ablation study validates the importance of each improvement point proposed in our method and further demonstrates its effectiveness on the OK-VQA dataset.

Table 2: The comparison of results from existing models on the A-OKVQA dataset.

Method	Direct Answer		Multiple Choice	
	Val	Test	Val	Test
ClipCap[19]	18.1	15.8	44.0	43.8
Pythia[36]	25.2	21.9	49.0	40.1
ViLBERT[28]	30.6	25.9	49.1	41.5
LXMERT[27]	30.7	25.9	51.4	41.6
KRISP[4]	33.7	27.1	51.9	42.2
GPV-2[37]	48.6	40.7	<u>60.3</u>	<u>53.7</u>
PromptCap[11]	56.3	<u>59.6</u>	73.2	73.1
Prophet[12]	58.2	55.7	59.3	57.3
SimpleBaseline[13]	<u>58.6</u>	57.5	-	-
MM-Reasoner[40]	-	60.2	-	-
Self-Bootstrapped[42]	57.2	56.4	-	-
ours	59.2	-	60.1	-

Table 3: The impact of different numbers of context examples n on the performance of the QA model when the number of query attempts $T=1$.

Model($T=1$)	Caption	Examples				
		n=0	n=1	n=8	n=16	n=20
GPT-3	VinVL	49.97	54.89	57.49	57.52	57.91
Mistral	VinVL	52.75	56.20	58.28	58.69	59.04
	PromptCap+VEGCAP	54.66	56.85	59.11	59.75	59.94

Tables 3 and 4 demonstrate that the proposed method can provide richer visual information to large language models, further enhancing their capabilities. The results in rows 1 and 2 indicate that, under the same conditions with a query count of $T=1$, the Mistral model outperforms GPT-3. However, at $T=5$, Mistral’s performance does not exceed that of GPT-3. Although five queries allow for the introduction of five times the number of different samples, the improvement for Mistral relative to GPT-3 is not significant. This may be because the Mistral model exhibits weaker learning capability when processing context examples that have a low similarity to the test sample, causing its generated answers to rely more on context examples that are similar to the test sample, thereby limiting performance enhancement. Nonetheless, both models exhibit

Table 4: The impact of different numbers of context examples n on the performance of the QA model when the number of query attempts $T=5$.

Model($T=5$)	Caption	Examples				
		n=0	n=1	n=8	n=16	n=20
GPT-3	VinVL	49.97	56.75	59.91	60.84	61.10
Mistral	VinVL	52.75	57.28	58.90	59.01	59.25
	PromptCap+VEGCAP	54.66	58.55	60.43	60.77	60.95

a consistent trend when faced with varying numbers of context examples: as the number of context examples N increases, VQA accuracy also improves. Notably, when $N=0$, the Mistral model performs zero-shot learning, answering questions based solely on the provided question, image caption, and candidate answers. The original Prophet method using GPT-3 achieved only 49.97% accuracy, which is far inferior to the performance of the MCAN model. Even when using the Mistral model for answers, its performance is still lower than that of the MCAN model. As indicated in Table 3, the proposed method demonstrates excellent performance in zero-shot learning scenarios, achieving an accuracy of 54.66% without the need for a large number of context examples, which is significantly higher than that of REACT, which also utilizes Mistral for zero-shot learning.

In Table 5 the first column presents the fine-tuning results of the original MCAN model and the proposed CapMCAN model on the OK-VQA dataset after pre-training on VQA 2.0. The second column shows the impact of the candidate answers and context examples generated by the two VQA models, using VinVL captions under the conditions of $N=20$ and $T=5$, on Mistral’s ability to answer questions. The results indicate that the CapMCAN model, with the introduction of visual entity-guided image captions, generates more accurate answers and further improves the quality of Mistral’s responses. The final results show a 0.3% increase in accuracy, validating the positive impact of visually grounded image descriptions on model performance.

Table 5: The impact of different VQA models on the Mistral model’s ability to answer questions.

Model	Finetuning	Mistral
MCAN	53.04	59.25
CapMCAN	53.42	59.56

Table 6 illustrates the impact of different image captions on the final results based on the candidate answers and context examples generated by the proposed CapMCAN model. From the results in rows 1 and 2, it can be observed that the accuracy using image captions generated by VinVL is 0.5% lower compared to using the general captions generated by OFA. This indicates that more powerful caption models can indeed enhance VQA accuracy. However, incorporating the general captions generated by the OFA model into the PromptCap captions does not significantly affect the experimental results, further validating the effectiveness of the proposed method. According to the experimental findings, the VEGCAP model’s captions provide only a 0.33% improvement compared to those generated by the VinVL model. This limited enhancement may be attributed to the fact that the information contained in the VEGCAP captions has already been utilized in the CapMCAN model. Furthermore, VEGCAP captions focus solely on the subjects of the questions without considering the context or relationships with surrounding objects. On the other hand, the PromptCap model considers the questions while generating captions, making them more oriented toward the entities in the questions. However, the captions provided are not sufficiently detailed. The VEGCAP model proposed in this study effectively addresses this limitation by offering richer and more contextualized captions.

Table 7 presents the performance of the VEGCAP model on various question types in the OK-VQA dataset. The model performs well on questions related to daily activities, particularly in the categories of Sports and Recreation (66.84%), Weather and Climate (64.50%), and Cooking and Food (63.03%), where the accuracy exceeds 60%. This indicates that the model is effective at handling questions related to common life scenarios and visual content. How-

Table 6: The impact of different captions on accuracy (acc) when $n=20$ and $T=5$.

VinVL	OFA	PromptCap	VEGCAP	Acc(%)
✓				59.56
	✓			60.06
		✓		60.53
	✓	✓		60.59
			✓	59.89
		✓	✓	60.95

ever, the model shows poorer performance on questions requiring more abstract reasoning or specialized knowledge, such as Science and Technology (50.00%) and Vehicles and Transportation (56.21%), where the accuracy is significantly lower than in other categories. This could be due to the complexity of the concepts or the rarity of the visual elements in these questions, as well as the model’s lack of sufficient domain-specific knowledge and reasoning capability. Furthermore, the accuracy in the Geography, History, Language and Culture (58.44%) and People and Everyday Life (59.02%) categories is moderate, suggesting that the model faces challenges when dealing with diverse questions that require a broader knowledge base. In summary, while the VEGCAP model performs well with questions that are closely related to daily life and visual information, it still shows notable weaknesses in tasks that require abstract reasoning and specialized domain knowledge. Future improvements should focus on enhancing domain-specific knowledge acquisition, as well as boosting the model’s performance in abstract reasoning and specialized fields, particularly in areas like science, technology, and complex visual scenarios, to better address the challenges posed by diverse and complex tasks.

Table 7: The accuracy on different types of questions in the OK-VQA dataset.

Question Type	acc(%)
Plants and Animals	62.93
Science and Technology	50.00
Sports and Recreation	66.84
Geography, History, Language and Culture	58.44
Brands, Companies and Products	58.37
Vehicles and Transportation	56.21
Cooking and Food	63.03
Weather and Climate	64.50
People and Everyday life	59.02
Objects, Material and Clothing	56.73
Overall	60.95

Figure 6 illustrates error cases in Mistral model responses after applying the method proposed in this paper. The first case on the left involves a question about the “brand of the suit is the man in the image wearing,” but the description only mentions “a man in a suit with a name badge on his chest,” lacking specific clues related to the brand. In the case on the right, the image caption includes the term “a statue of buddha,” yet the model still incorrectly responds

with "chinese." The model might have mistakenly established a simple association with Chinese culture based on the combination of Buddha statues and pandas, ignoring the religious context specified in the question. In the bottom-left case, the image description mentions "an old car," but the model chooses "antique" as the answer, likely reflecting its insufficient ability to match the car type with the description. Although the bottom-right case is classified as incorrect, from a human perspective, the model's answer is actually correct. This highlights that over-reliance on precise matching evaluation may lead to misjudgments when dealing with rich and vivid answers, necessitating a more scientifically grounded evaluation method to better assess the model's performance.

To optimize the model's response accuracy, it is essential to improve the diversity and quality of training data, ensuring that it covers more varied scenes and detailed features, which helps enhance the model's ability to handle different situations. Additionally, introducing multimodal alignment optimization mechanisms can further improve the model's ability to integrate visual and linguistic information, leading to more accurate answer generation. These optimization measures will significantly enhance the model's capacity to respond to complex queries, reduce misjudgments, and improve overall performance.



Figure 6: The cases where the proposed method in this paper gives incorrect answers on the OK-VQA dataset, with Q, GT, C, CA, and A representing the Question, Ground Truth, Caption, Candidate Answers with their Confidence Scores, and the Answer predicted by the model, respectively.

Figure 7 show cases the effectiveness of the proposed method, demonstrating how the model utilizes image captions or candidate answers to respond to questions in the given cases. In the upper right case, the model answers the question, 'Why are the boys wearing the same shirts?' by leveraging information from the image caption, such as 'a boy kicking a soccer ball' and 'green

VEGCAP: A Visual Entity-Guided Image Caption Generation method for Visual Question Answer



Q: What is the term for the kind of furniture in front of the mirror?

GT: ['ottoman', 'ottoman', 'ottoman', 'ottoman', 'ottoman', 'ottoman', 'foot rest', 'foot rest', 'futon', 'futon']

C: a room with a mirror and an ottoman; a ottoman sitting on the floor in a room.

CA: chair (0.40); sofa(0.09); couch(0.04); loveseat(0.01); bench(0.01); ottoman(0.01); seat(0.01); fold(0.01); sit(0.00); lounge(0.00)

A: ottoman

Acc: 1.0



Q: What are the white objects on this animals head called?

GT: ['horn', 'horn', 'horn', 'horn', 'horn', 'horn', 'horn', 'horn', 'tusk', 'tusk']

C: a brown cow laying on a grassy field; a brown bull with horns standing in a field.

CA: horn (0.40); tusk(0.15); horns(0.01); ivory(0.00); cow(0.00); fur(0.00); nylon(0.00); banana(0.00); mix(0.00); cattle(0.00)

A: horn

Acc: 1.0



Q: Why do the boys shirts match?

GT: ['same team', 'same team', 'same team', 'same team', 'uniform', 'uniform', 'team', 'team', 'team mate', 'team mate']

C: a boy kicking a soccer ball on a field; a young boy in a green and white soccer uniform running.

CA: team (0.10); sport(0.10); uniform(0.04); color(0.01); athletic(0.01); snow(0.00); wed(0.00); fun(0.00); military(0.00); beer(0.00)

A: team

Acc: 0.6



Q: What type of process is occurring here?

GT: ['make donuts', 'make donuts', 'make donuts', 'make donuts', 'fry', 'fry', 'donut fry', 'donut fry', 'doughnut make', 'doughnut make']

C: a donut maker making donuts; a bunch of donuts being fried in a machine

CA: eat (0.08); cook(0.05); bake(0.04); fry(0.01); roast(0.01); age(0.01); measure(0.01); birthday(0.00); donuts(0.00); party(0.00)

A: fry

Acc: 0.6

Figure 7: The cases where the proposed method in this paper gives correct answers on the OK-VQA dataset.

and white soccer uniform'. The model not only relies on the image caption but also draws on its external knowledge to infer that the boys are likely members of the same soccer team, which explains their identical uniforms. By integrating its understanding of sports and team uniforms, the model gains a deeper insight into the context of the question, enabling it to make a reasonable deduction. Additionally, it is noteworthy that the confidence score for the correct answer in the candidate answers may not necessarily be the highest. As illustrated in Figure 6, the correct answer might not even be present among the candidate answers. In such instances, the model must synthesize information from available image captions, candidate answers, and contextual examples to deduce the correct answer to the question. This demonstrates that the model possesses reasoning abilities even when direct correct answers are absent, thereby enhancing the accuracy of its responses.

5. Conclusion

In this paper, we proposed an innovative image caption generation strategy for visual question answering (VQA) tasks. This strategy guides the model to focus on and describe specific regions of an image based on a natural language question. Specifically, the model generates captions of the image regions relevant to the question and uses these captions, along with the question and candidate answers, as input to a language model to solve the visual question answering task. Moreover, our improved CapMCAN model effectively leverages the visual information contained in the image captions, further enhancing the accuracy of the model's responses. Through a series of extensive ablation studies and comparative experiments, our results demonstrate that

our method outperforms most existing approaches in terms of accuracy and performance. Additionally, this approach is adaptable to more powerful models, suggesting significant potential for future applications.

However, this approach of relying on image captions to convey image information has notable limitations. First, image descriptions may omit crucial information or generate incorrect representations, directly impacting the quality of the model's responses. Moreover, the performance of the model on domain-specific questions is often suboptimal, as it lacks the necessary specialized knowledge. Additionally, the model's ability to fuse multimodal information is limited, leading to less accurate cross-modal alignment between images and text, which constrains its reasoning and comprehension. Future improvements should focus on integrating domain-specific knowledge bases and enhancing the alignment and integration of image-text modalities to boost the model's performance in complex and specialized domains.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62166050), Major Science and Technology Project of Yunnan Province (202402AD080002), Key Program of Fundamental Research Project of Yunnan Science and Technology Plan (202201AS070021), and Scientific research foundation of Yunnan Provincial Department of Education (2024Y159)

References

- [1] Antol S, Agrawal A, Lu J, et al.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp. 2425-2433 (2015)
- [2] Marino K, Rastegari M, Farhadi A, et al.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp. 3195-3204 (2019)
- [3] Gardères F, Ziaefard M, Abeloos B, et al.: Conceptbert: Concept-aware representation for visual question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 489-498 (2020)
- [4] Marino K, Chen X, Parikh D, et al.: Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14111-14121 (2021)
- [5] Wu J, Lu J, Sabharwal A, et al.: Multi-modal answer validation for knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence, 36(3), pp. 2712-2721 (2022)
- [6] Gao F, Ping Q, Thattai G, et al.: Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5067-5077 (2022)
- [7] Luo M, Zeng Y, Banerjee P, et al.: Weakly-supervised visual-retriever-reader for knowledge-based question answering. arxiv preprint arxiv:2109.04014 (2021)
- [8] Lin W, Byrne B.: Retrieval augmented visual question answering with outside knowledge. arxiv preprint arxiv:2210.03809 (2022)
- [9] Lin W, Chen J, Mei J, et al.: Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. Advances in Neural Information Processing Systems. 36, 22820-22840 (2023)
- [10] Yang Z, Gan Z, Wang J, et al.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence, 36(3), pp. 3081-3089 (2022)
- [11] Hu Y, Hua H, Yang Z, et al.: Promptcap: Prompt-guided task-aware image captioning. arxiv preprint arxiv:2211.09699 (2022)
- [12] Shao Z, Yu Z, Wang M, et al.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 14974-14983 (2023)
- [13] Xenos A, Stafylakis T, Patras I, et al.: A simple baseline for knowledge-based visual question answering. arxiv preprint arxiv:2310.13570 (2023)

- [14] Barezi E J, Kordjamshidi P.: Disentangling Knowledge-based and Visual Reasoning by Question Decomposition in KB-VQA. arxiv preprint arxiv:2406.18839 (2024)
- [15] Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. *Advances in neural information processing systems*. 33, 1877-1901 (2020)
- [16] Jiang A Q, Sablayrolles A, Mensch A, et al.: Mistral 7B. arxiv preprint arxiv:2310.06825 (2023)
- [17] Touvron H, Martin L, Stone K, et al.: Llama 2: Open foundation and fine-tuned chat models. arxiv preprint arxiv:2307.09288 (2023)
- [18] Chen S, Wang W, Chen X, et al.: Llama-lora neural prompt engineering: A deep tuning framework for automatically generating chinese text logical reasoning thinking chains. *Data Intelligence*. 1-53 (2024)
- [19] Schwenk D, Khandelwal A, Clark C, et al.: A-okvqa: A benchmark for visual question answering using world knowledge. In: *European conference on computer vision*. Cham: Springer Nature Switzerland, pp. 146-162 (2022)
- [20] Speer R, Chin J, Havasi C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: *Proceedings of the AAAI conference on artificial intelligence*, 31(1), pp. (2022)
- [21] Vrandečić D, Krötzsch M.: Wikidata: a free collaborative knowledgebase. *Communications of the ACM*. 57(10), 78-85 (2014)
- [22] Gui L, Wang B, Huang Q, et al.: Kat: A knowledge augmented transformer for vision-and-language. arxiv preprint arxiv:2112.08614 (2021)
- [23] Lin Y, Xie Y, Chen D, et al.: Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*. 35, 10560-10571 (2022)
- [24] Li X, Yin X, Li C, et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer International Publishing, pp. 121-137 (2020)
- [25] Zhang P, Li X, Hu X, et al.: Vinvl: Revisiting visual representations in vision-language models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579-5588 (2021)
- [26] Su W, Zhu X, Cao Y, et al.: Pre-training of generic visual-linguistic representations. arxiv preprint arxiv:1908.08530 (2019)
- [27] Tan H, Bansal M.: Lxmert: Learning cross-modality encoder representations from transformers. arxiv preprint arxiv:1908.07490 (2019)
- [28] Lu J, Batra D, Parikh D, et al.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*. 32, (2019)
- [29] Alayrac J B, Donahue J, Luc P, et al.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*. 35, 23716-23736 (2022)
- [30] Li J, Li D, Savarese S, et al.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*, PMLR, pp. 19730-19742 (2023)
- [31] Wang P, Yang A, Men R, et al.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: *International conference on machine learning*, PMLR, pp. 23318-23340 (2022)
- [32] Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need. *Advances in neural information processing systems*. 30, (2017)
- [33] Yu Z, Yu J, Cui Y, et al.: Deep modular co-attention networks for visual question answering. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6281-6290 (2019)
- [34] Li Y, Chen X, Hu B, et al.: Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. arxiv preprint arxiv:2402.13561 (2024)
- [35] Liang M, Wu Y: TOA: task-oriented active VQA. *Advances in Neural Information Processing Systems*. 36, (2024)
- [36] Jiang Y, Natarajan V, Chen X, et al.: Pythia v0. 1: the winning entry to the vqa challenge 2018. arxiv preprint arxiv:1807.09956 (2018)
- [37] Kamath A, Clark C, Gupta T, et al.: Webly supervised concept expansion for general purpose vision models. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, pp. 662-681 (2022)
- [38] Devlin J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv preprint arxiv:1810.04805 (2018)
- [39] Rao J, Shan Z, Liu L, et al.: Retrieval-based knowledge augmented vision language pre-training. In: *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5399-5409 (2023)
- [40] Khademi M, Yang Z, Frujeri F, et al.: MM-Reasoner: A Multi-Modal Knowledge-Aware Framework for Knowledge-Based Visual Question Answering. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6571-6581 (2023)
- [41] Li M, Li H, Du Z, et al.: Rationale-based Ensemble of Multiple QA Strategies for Zero-shot Knowledge-based VQA. arxiv preprint arxiv:2406.12746 (2024)
- [42] Hao D, Wang Q, Guo L, Jiang J, Liu J.: Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering. arxiv preprint arxiv:2404.13947 (2024)