Diversify, Rationalize, and Combine: Ensembling Multiple QA Strategies for Zero-shot Knowledge-based VQA

Miaoyu Li, Haoxin Li, Zilin Du, and Boyang Li
College of Computing and Data Science,
Nanyang Technological University, Singapore
{miaoyu.li, haoxin003, zilin003, boyang.li}@ntu.edu.sg

Abstract

Knowledge-based Visual Qustion-answering (K-VQA) often requires the use of background knowledge beyond the image. However, we discover that a single knowledge generation strategy is often insufficient for all K-VQA questions. To this end, we propose **Di**versification, **E**vidence **T**runcation, and **Co**mbination for **K**nowledge-based **E**lucidation (DIETCOKE), which utilizes a bundle of complementary question-answering tactics and aggregates their answers using textual rationales. DIETCOKE comprises of three stages: diversification, rationalization, and ensemble. The diversification stage generates three distinctive decision contexts, each leading to its own answer candidate. The rationalization stage generates two rationales, the automatic rationale and the mechanistic rationale, for each answer candidate using decorrelated techniques. Finally, in the ensemble stage, an LLM informed by the rationales selects one answer from the three candidates. Experiments show that DIETCOKE significantly outperforms state-of-the-art LLM-based baselines by 2.8% on OK-VQA and 4.7% on A-OKVQA and that the strategies in the ensembles are highly complementary. Code is available at: https://github.com/limiaoyu/DietCoke

1 Introduction

Knowledge-based Visual Question-answering (K-VQA) requires background knowledge beyond the image content. For example, to answer the question in Fig. 1 (b), it is necessary to know the carbohydrate content of different types of food. Zero-shot K-VQA provides an effective evaluation of AI models in applying their knowledge to answer novel vision-informed questions.

An effective technique for zero-shot VQA is to translate an image to a textual decision context for a text-based model, such as an LLM, which the answers the question using the context (Yang et al., 2022; Tiong et al., 2022). For example, Yang et al. (2022) translate the image to captions; Guo et al. (2023) further generate question-answer pairs to demonstrate the VQA task. To cater to the knowledge-intensive nature of K-VQA, Cao and Jiang (2024) generate a paragraph describing relevant background knowledge, from which an LLM extracts the answer. These approaches rely on a single question-answering strategy using a single type of decision context.

Interestingly, we observe that one LLM strategy is often not sufficient for K-VQA datasets. For some questions, merely the captions are sufficient for finding the right answer. Other questions benefit from knowledge extracted from LLMs, but it is often not clear if a short, one-sentence knowledge statement or a paragraph would beget the right answer from the LLM. As an illustration, in Fig. 1 we show three questions and three AI-generated decision contexts for each question, including captions, short-form knowledge as a single sentence, and long-form knowledge, which contains multiple sentences. In each question, only one decision context leads to the right answer.

In light of this, we propose **Di**versification, **E**vidence **T**runcation, and **Co**mbination for **K**nowledge-based **E**lucidation (DIETCOKE) to achieve dynamic ensemble of different question-answering strategies. DIETCOKE comprises a diversification phase, a rationalization phase, and an ensemble phase. In each phase, we use an in-context-learning LLM without training. In diversification, we generate three decision contexts for each K-VQA question, including image captions, short-form knowledge, and long-form knowledge, with an increasing amount of background knowledge. The LLM generates three answer candidates, one from each decision context.

In the ensemble phase, the LLM selects one correct answer from the three candidates. To encourage informed decisions, we design the rationalization phase, which generates two complementary

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1552–1566
November 12-16, 2024 ©2024 Association for Computational Linguistics

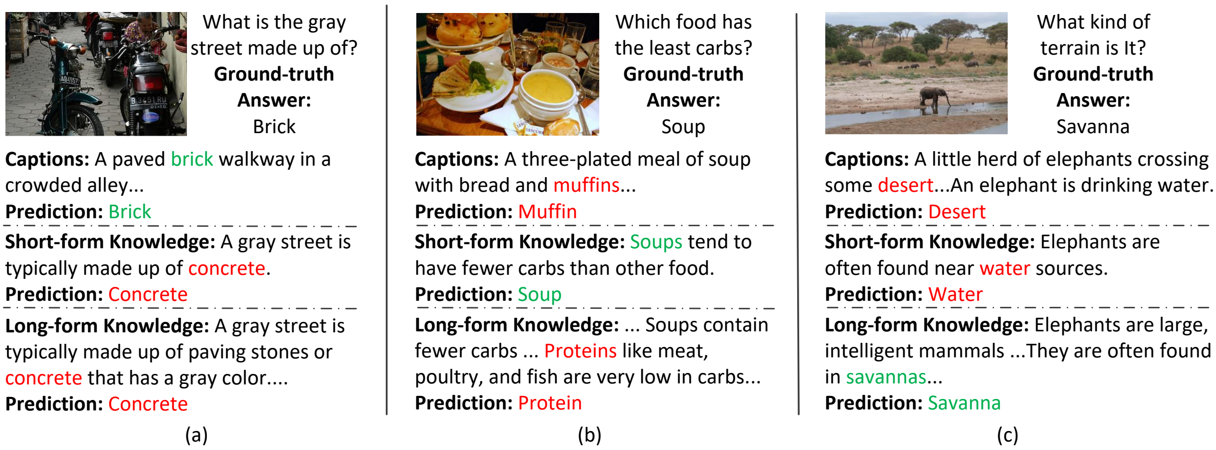

Figure 1: The three K-VQA questions are best answered using distinctive decision contexts, including image captions, (a), and two forms of generated knowledge statements, (b) and (c). This is due to (1) the difficulty in controlling the generation of captions and knowledge statements and ensuring they contain relevant information, and (2) the inability of LLMs to identify correct answers from noisy contexts.

types of rationales for each answer candidate. A rationale summarizes the portion of a decision context that can support the correctness of the answer. The first type, automatic rationale, is generated by simply asking an LLM to summarize the reasoning behind the answer into one sentence. The second type, mechanistic rationale, is one sentence from the decision context that contributes the most to the generation of the answer, as identified by a model interpretation technique, GradCam (Selvaraju et al., 2017). These brief rationales preserve the most relevant information in the original decision contexts, and prevent irrelevant information from misleading the ensemble LLM. Informed by the rationales, the LLM in the ensemble phase selects one final answer from the candidates, achieving dynamic ensemble of the question-answering strategies.

On two popular K-VQA datasets, OKVQA and A-OKVQA, DIETCOKE outperforms state-of-the-art baselines that use frozen LLMs without training by 2.6% to 4.7%. Ablation studies reveal that the fusion of three answers improves performance over the best single answer strategy by 1.1% on OK-VQA and 1.6% on A-OKVQA. Further, adding both the automatic and mechanistic rationales obtains gains of 1.2% and 1.4%, respectively on the two datasets.

The success of DIETCOKE can be understood from multiple perspectives. First, DIETCOKE creates an effective ensemble by employing decorrelated QA strategies and rationalization strategies. Classic theory on ensemble learning (Clemen and Winkler, 1985; Breiman, 2001) indicates that an ensemble of weak classifiers becomes more powerful as the individual classifiers become less correlated and more diverse. Intuitively, different QA contexts lead to diverse answers. Also, disparate mechanisms used by the two rationalization strategies utilize reduce correlation. In §4.3, we perform a series of ablation studies that demonstrate the synergy among the three QA strategies and the two rationalization strategies.

Second, rationales in DIETCOKE can be understood from a chain-of-thought (CoT) perspective. Recently, Li et al. (2024) prove that chain-of-thought prompting expands the circuit complexity of problems solvable by an LLM from the class AC^0 with constant circuit depth to the class $P/poly$ with polynomial circuit depth. This is made possible by saving intermediate computational results as CoT tokens. Analogically, providing the QA contexts (*i.e.,* intermediate results) to the answer-selecting ensemble LLM should be beneficial. However, in practice we observe that long decision contexts may contain many irrelevant facts that can mislead the ensemble LLM. By summarizing the contexts into rationales, we provide abridged chains-of-thoughts that expand circuit depth while avoiding misleading tokens.

Our contributions include:

- We identify the need for combined use of multiple question-answering strategies in K-VQA, and propose DIETCOKE, which fuses multiple answers and answer strategies dynamically using frozen in-context learning LLMs, achieving state-of-the-art performance on zero-shot

OKVQA and A-OKVQA.

- We propose that rationales serve an important role in answer ensemble. Further, we devise two rationale generation techniques, automatic LLM summarization, and mechanistic LLM interpretation. These rationales complement each other in enhancing VQA performance.

2 Related Work

Knowledge-based VQA. Knowledge-based VQA requires background knowledge beyond the image content to answer questions. The popular K-VQA datasets include OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), both of which cover a wide range of knowledge categories. For OK-VQA, answering questions usually only requires knowledge recall, without the need for additional reasoning. A-OKVQA, on the other hand, is an enhanced version of OK-VQA, where questions often require further reasoning based on background knowledge to answer. Early studies (Gardères et al., 2020; Marino et al., 2021; Wu et al., 2022; Zhu et al., 2020; Narasimhan et al., 2018; Narasimhan and Schwing, 2018; Gao et al., 2022) retrieve related knowledge from external knowledge bases such as Wikipedia, ConceptNet, and Google Images. Recent approaches (Yang et al., 2022; Hu et al., 2022; Wang et al., 2023; Shao et al., 2023; Xing et al., 2023; Si et al., 2023; Khan et al., 2024) require no explicit knowledge retrieval; they leverage implicit world knowledge stored in LLMs or Large Vision-Language Models (LVLMs) by using them as QA models. To further improve the QA performance of LLMs or LVLMs, many few-shot methods (Yang et al., 2022; Shao et al., 2023; Xenos et al., 2023; Khan et al., 2024) are proposed, which use the training samples as exemplars in the prompt. Unlike existing LLM-based methods that utilize knowledge implicitly, our method explicitly generates relevant background knowledge from an LLM and uses it to assist in answering questions.

End-to-end LVLMs and LVLM Prompting for Zero-shot K-VQA. End-to-end training is adopted to align the vision and text modalities, so that a pretrained visual encoder can work seamlessly with a pretrained LLM, resulting in LVLMs (Wang et al., 2022; Alayrac et al., 2022; Liang et al., 2022; Mañas et al., 2022; Li et al., 2023; Jiang and Zheng, 2023; Zhu et al., 2023; Lin et al., 2024; Liu et al., 2024). While these LVLMs achieve high performance on zero-shot K-VQA, they require training on an enormous quantity of image-text pairs and are hence compute-intensive.

To enable the LVLMs to better understand questions and perform reasoning, one concurrent method explore question rephrasing in the prompt (Prasad et al., 2024). However, it is difficult for this method to change to a different LLM, which provides crucial knowledge for K-VQA, without expensive retraining of the LVLMs. In contrast, our method translates images into textual descriptions and directly utilizes frozen LLMs to answer questions, enabling us to easily change the knowledge source at any time.

Text-mediated Zero-shot K-VQA. Another type of zero-shot method converts visual information to textual descriptions and applies frozen LLMs to answer questions based on these descriptions alone. We classify these methods based on the number of types of descriptions used and whether answer ensemble is employed. Most methods (Yang et al., 2022; Tiong et al., 2022; Guo et al., 2023; Chen et al., 2024) translate images into a single type of description: captions, and generate one answer from them. Chen et al. (2024) uses two different LVLMs to generate different captions and answer candidates, and one LLM to choose the final answer. To accommodate the need for external knowledge in K-VQA, Cao and Jiang (2024) generate two types of descriptions, captions and long-form knowledge and concatenates into one prompt. Hence, we consider it to be a single-strategy method. In contrast, DIETCOKE explicitly combines different QA strategies by selecting from their answers.

3 Method

The goal of our method is to jointly utilize multiple decision contexts to answer each question and achieve a dynamic ensemble of these different strategies. To accomplish this, we propose DIETCOKE, which consists of three phases: diversification, rationalization and ensemble. This section provides detailed descriptions of them.

3.1 Diversification

In diversification phase, we generate three types of decision contexts for each question, including captions, short-form knowledge consisting of one sen-

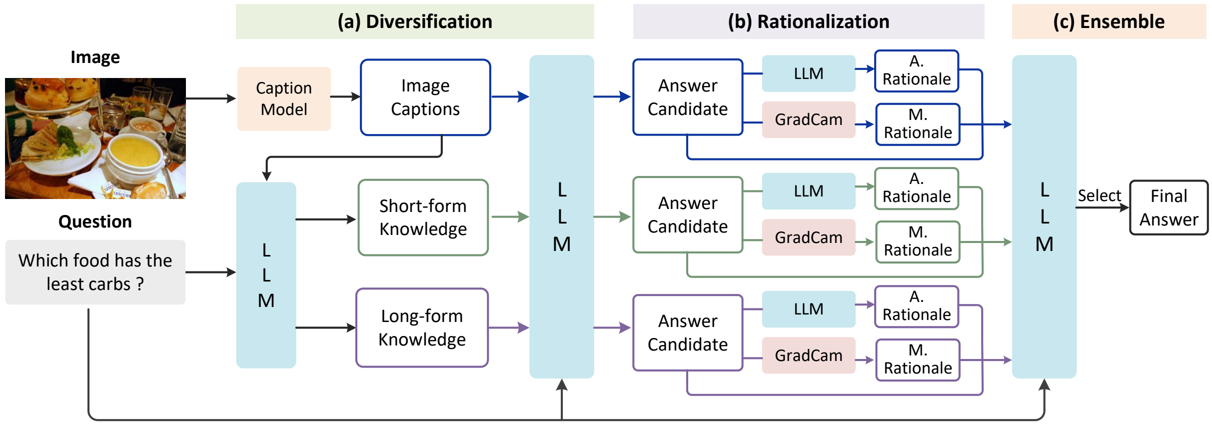

Figure 2: The framework of DIETCOKE. The diversification phase contains three question-answering strategies, which generate three different decision contexts for the answer-predicting LLM. The short-form knowledge contains a single sentence, whereas the long-form knowledge consists of a paragraph of background information. The caption-only strategy generates only image captions and no background knowledge. For each strategy, the LLM generates an answer candidate. In the rationalization phase, for each candidate, we generate an Automatic (A.) Rationale and a Mechanistic (M.) Rationale, which guide the ensemble LLM in selecting the best answer.

tence, and long-form knowledge containing multiple sentences. They provide three different strategies for each question, resulting in three answer candidates.

The Caption-only Strategy. Since LLMs can only process textual information, we use an off-the-shelf image captioning model to transform an image into multiple textual captions. To ensure the relevance to the question, we employ the question-guided caption generation mechanism in PNP-VQA (Tiong et al., 2022), which first identifies the image patches most relevant to the question and then generates captions only from these relevant patches. The prompt for the caption-only strategy is shown in Tab. 4 of Appendix B.

To achieve in-context learning, we follow Img2LLM (Guo et al., 2023) and generate Question-answer (QA) pairs from the image captions. The QA pairs serve as demonstration of the QA task for LLM in-context learning. As the task demonstration is necessary for LLMs to perform QA, we include these QA pairs in the prompt of *all three* strategies of diversification.

The Short-form Knowledge Strategy. We generate relevant background knowledge for each question using an LLM. To ensure that the generated knowledge is relevant to the question and the image content, we provide the captions and the question in prompt: "$\langle$*Captions*$\rangle$ $\langle$*Question*$\rangle$ *Please provide background knowledge related to this question in a single sentence.*". To generate short-form knowledge, we add the length constraint "*in a single sentence*". The short-form knowledge ideally represents the most relevant background knowledge to the question. The prompts for generating short-form knowledge and the answer to the question are detailed in Tab. 5 and Tab. 4 in Appendix B.

We observe that short-form knowledge tends to be highly relevant to the visual question. However, due to its short length, it may not be comprehensive enough to get the correct answer.

The Long-form Knowledge Strategy. In this strategy, we remove the length constraint "*in a single sentence*" from the prompt, so that the LLM can generate as much text as it wants. In most cases, this leads to one or two paragraphs, which ideally could capture comprehensive background knowledge related to the question. The prompts for generating long-form knowledge and utilizing it to generate answer are detailed in Tab. 5 and Tab. 4 of Appendix B.

The short-form knowledge and the long-form knowledge are synergistic. Constrained to a single sentence, the short-form knowledge context is succinct but may leave out relevant information. In comparison, the long-form knowledge strategy prioritizes knowledge recall over precision, and may contain many irrelevant facts that potentially distract the LLM. Their synergy is further demonstrated in the experiments (§4.3 and Fig. 4). Balancing different strategies is hence important for high performance.

3.2 Rationalization

Since not all question-answering strategies are suitable for a given question, selecting the correct answer is crucial. To allow the LLM to make informed choices, the rationalization phase generates two types of rationales with different methods for each answer candidate, which we call automatic rationales and mechanistic rationales. These brief rationales retain only the most relevant information from the decision context, thus mitigating the risk of the LLM being misled by a plethora of irrelevant information.

Automatic Rationale. To generate the automatic rationale, we feed the original question, the decision context, and the predicted answer to an LLM and directly ask the LLM to summarize the rationale behind the answer to a single sentence. The LLM is not restricted to selecting a sentence from the decision context and performs open-ended generation. The prompt for generating automatic rationales is shown in Tab. 6 of Appendix B.

Mechanistic Rationale. Though the automatic rationale is easy to acquire and often reasonable, there are occasions when the LLM generate incorrect rationales by ignoring part of the decision context or hallucinating (see examples in Fig. 8). Although this could be partially alleviated by exhaustive prompt engineering, we opt for a more systematic solution, which is to obtain the rationale through mechanistic interpretation of the answer-generating LLM.

Inspired by GradCAM (Selvaraju et al., 2017), we devise a method to compute the contribution of each sentence in the decision context to the answer candidate. We first introduce some notations. The prompt to the LLM, including the instruction and the decision context (captions or knowledge statements), contains N_P tokens. The answer generated by the LLM contains N_A tokens. When generating the k-th answer token w_k, the Transformer-based LLM attends to all previous $N_k = N_P + k - 1$ tokens. We extract the attention weight vector $\boldsymbol{a}^{(h)} \in \mathbb{R}^{N_k}$ from the h-th attention head. The components of $\boldsymbol{a}^{(h)}$ sum up to 1, $\sum_{i=1}^{N_k} a_i^{(h)} = 1$.

We seek the contribution of every token preceding the k-th answer token. Instead of the raw attention scores, which may be redundant and inaccurate, we weigh the attention scores with its gradient from the probability p_k of the predicted answer token w_k. We disregard negative gradients as we focus on positive contributions. Formally, the relevance scores of each of the N_k tokens is

$$\boldsymbol{r}_k = \frac{1}{H}\sum_{h=1}^{H} \max\left(0, \frac{\partial p_k}{\partial \boldsymbol{a}^{(h)}}\right) \boldsymbol{a}^{(h)}, \qquad (1)$$

where H is the number of attention heads. Next, we re-normalize and aggregate the contribution of prompt tokens to all answer tokens.

$$\boldsymbol{r} = \sum_{k=1}^{N_A} \mathit{softmax}(\boldsymbol{r}_k[0 : N_P]), \qquad (2)$$

where $[0 : N_P]$ selects the first N_P components of $\boldsymbol{r}_k$. Finally, we obtain sentence-level contributions by summing over the contributions of all tokens in each sentence in the decision context. The sentence with the highest contribution is picked as the mechanistic rationale. As the short-form knowledge context contains only a single sentence, it is always selected as the mechanistic rationale.

3.3 Ensemble

In the final ensemble stage, we provide all three answer candidates with corresponding rationales, including both the automatic and the mechanistic rationales, in the prompt. The LLM is instructed to select the best answer, thereby achieving dynamic strategy ensemble. To ensure the model perceives the image content during ensemble, we also include caption-generated QA pairs in the prompt. The exact prompt for question-answering strategy fusion is in Tab. 7 of Appendix B.

4 Experiment

4.1 Setup

Datasets. We evaluate our method on two mainstream K-VQA datasets: OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022). Questions in both datasets require knowledge beyond the images to answer. We utilize the test split of OK-VQA and the validation split and the test split of A-OKVQA for evaluation. These splits contain 5,046, 1,100, and 6,700 questions, respectively. We follow the official evaluation protocols of direct answer and report VQA scores for each dataset.

Implementation Details. Since the quality of captions significantly impacts the results, we follow the previous approaches (Guo et al., 2023; Lan et al., 2023; Cao and Jiang, 2024) to use BLIP (Li

Method	QA Model	Model Size	Decision Context Captions	Decision Context Knowledge	OK-VQA test	A-OKVQA val	A-OKVQA test
LLM-based Zero-shot Methods							
PICa	GPT-3	175B	✓	×	17.7	-	-
PNP-VQA	UnifiedQA	11B	✓	×	35.9	-	-
Cola-Zero	FlanT5	11B	✓	×	39.4	-	-
Img2LLM	OPT	175B	✓	×	45.6	42.9	40.7
Img2LLM*	Gemma	7B	✓	×	45.6*	44.9*	-
Img2LLM*	Mistral	7B	✓	×	46.3*	44.3*	-
KGenVQA	UnifiedQA	11B	✓	long	45.4	39.1	-
RQprompt	GPT-3	175B	✓	×	46.4	43.2	43.9
DIETCOKE	Gemma	7B	✓	long & short	<u>47.6</u>	47.3	<u>46.8</u>
DIETCOKE	Mistral	7B	✓	long & short	**49.2**	<u>47.5</u>	**48.6**
LVLM-based Zero-shot Methods							
REPARE	BLIP2	3B	-	-	-	44.9	-
REPARE	BLIP2	11B	-	-	-	47.3	-
REPARE	MiniGPT-4	7B	-	-	-	33.2	-
REPARE	MiniGPT-4	13B	-	-	-	**47.9**	-

Table 1: Comparison with state-of-the-art methods on zero-shot K-VQA. We report the VQA score of direct answer on OK-VQA and A-OKVQA datasets. The best score is indicated in **bold**, while the second best score is indicated in <u>underline</u>. The results marked with * represent the baselines we implemented ourselves.

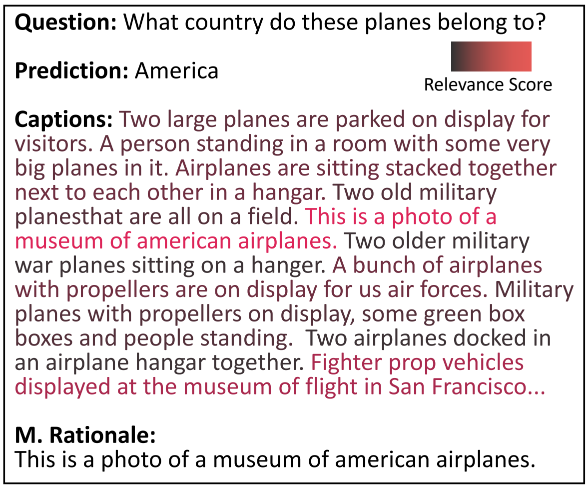

Figure 3: The heatmap of relevance scores of sentences in captions. The relevance score represents the contribution of the sentence to the answer.

et al., 2022) as the captioning model. We generate 30 captions for each image. Additionally, we utilize a finetuned T5-large model (Raffel et al., 2020), as in Img2LLM, to generate 30 QA pairs from the captions to enable in-context learning when generating answers. To demonstrate the generalization ability of our method, we use two different models, Mistral-7B and Gemma-7B, in our experiments. The LLMs cannot access answer lists or training samples, achieving zero-shot K-VQA.

Baselines. We compare DIETCOKE with previous zero-shot K-VQA methods without training. The methods using frozen LLMs can be roughly divided into two categories: (1) Methods employing the caption-only strategy, PICa (Yang et al., 2022), PNP-VQA (Tiong et al., 2022), Cola-Zero (Chen et al., 2024), Img2LLM (Guo et al., 2023), and RQprompt (Lan et al., 2023). (2) Methods employing the long-form knowledge strategy. The only method in this category is KGenVQA (Cao and Jiang, 2024). Unlike translating visual information into textual descriptions for frozen LLMs, the concurrent method REPARE (Prasad et al., 2024) directly utilizes an LVLM as QA model[1].

4.2 Main Results

To demonstrate the effectiveness of our dynamic ensemble of question-answering strategies, we compare our method with state-of-the-art baselines using different types of decision contexts and frozen LLMs. The results are shown in Tab. 1.

DIETCOKE establishes a new state-of-the-art on LLM-based zero-shot KVQA. Compared to previous best scores achieved by frozen LLMs, DIETCOKE with Mistral-7B outperforms them with large margins of 2.8%, 2.6% and 4.7% on the OK-

[1]The REPARE+LLaVA-1.5 model does not meet the zero-shot requirement, as LLaVA-1.5 is pretrained on OK-VQA and A-OKVQA. Hence, we omit that result from the comparison.

	Cap.	SK	LK	OK-VQA test	A-OKVQA val
#1	✓			46.3	44.3
#2		✓		48.1	45.9
#3			✓	46.7	45.5
#4	✓	✓		<u>48.4</u>	<u>47.1</u>
#5	✓		✓	47.1	46.8
#6	✓	✓	✓	**49.2**	**47.5**

Table 2: Ablation study results of diversification with Mistral-7B. Cap. denotes captions. SK and LK are short-form knowledge and long-form knowledge, respectively.

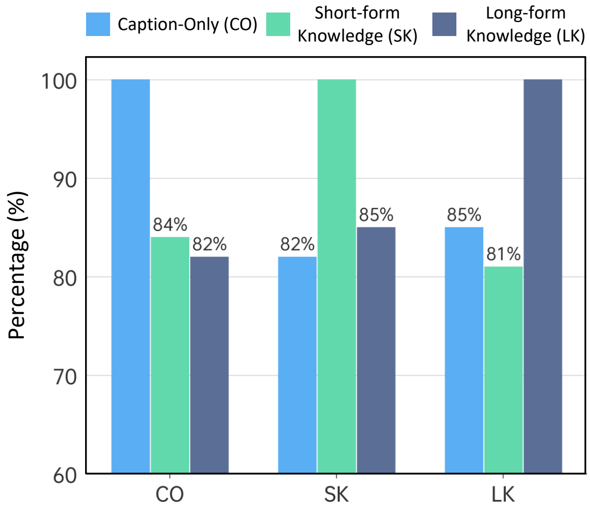

Figure 4: The success rate of the other two strategies on all questions successfully answered by one strategy. In left group, we show the percentage of correctly answered questions by SK and LK out of all the questions that are correctly answered by CO. The other two bar groups are drawn similarly.

VQA, A-OKVQA validation split and test split, respectively. These results strongly demonstrate the effectiveness of our approach.

We also compare against a concurrent method, REPARE (Prasad et al., 2024), which directly prompts an end-to-end trained LVLM. Since REPARE does not translate image to text, it avoids any information loss in the process. Nevertheless, DIETCOKE remains highly competitive, outperforming all results of REPARE but that from MiniGPT-4-13B with a much larger LLM, and the difference is merely 0.4%. The fact that a blind LLM, when prompted properly, is competitive with a well-prompted LVLM at a visual task is surprising to us. As the text-as-visual-representation approach can benefit from advances in LLMs without extensive vision-language alignment training, DIETCOKE offers strong practical benefits.

We also compare the time cost of DIETCOKE with Img2LLM on Mistral-7B. We randomly select 50 questions from the OKVQA test split and 50 from the A-OKVQA val split. We report the average time spent on each question using two RTX 3090 24G GPUs. Img2LLM takes 1.7 seconds. If we maximize parallelism of LLM calls, DIETCOKE takes 15.8 seconds. In completely sequential execution, DIETCOKE takes 29.5 seconds. DIETCOKE provides another point on the time-accuracy Pareto front and allows a VQA system to trade compute for answer quality when additional compute is available.

4.3 Ablation Studies

The QA Strategies. We systematically ablate the three answer strategies, the caption-only, the short-form knowledge, and the long-form knowledge. The outcomes are detailed in Tab. 2. Ablations #1, #2, and #3 utilize each question-answering strategy independently. Ablations #4 and #5 utilize partial combinations of the strategies.

	MR	AR	OKVQA test	A-OKVQA val
#1	Rand. Answer Selection		47.4	45.0
#2			48.0	46.1
#3	✓		48.5	46.4
#4		✓	<u>48.7</u>	<u>47.2</u>
#5	Rand.	✓	48.7	46.6
#6	✓	✓	**49.2**	**47.5**

Table 3: Ablation study of rationalization with Mistral-7B. MR and AR denote mechanistic rationale and automatic rationale, respectively.

Comparing #1 with #2 and #3, we observe that the short-form knowledge strategy and long-form knowledge strategy achieve better performance than caption-only strategy, demonstrating that generated background knowledge is advantageous in K-VQA. After fusing caption-only and short-form knowledge in #4, the performance on OKVQA and A-OKVQA improves by 2.1% and 2.8% over #1. This demonstrates the importance of strategy ensemble and the effectiveness of the short-form knowledge strategy. The same phenomenon can be observed in #1 and #5. Finally, the full DIETCOKE system, #6, achieves the best performance.

Further, we quantify how much each QA strategy complements others. In Fig. 4, for each QA

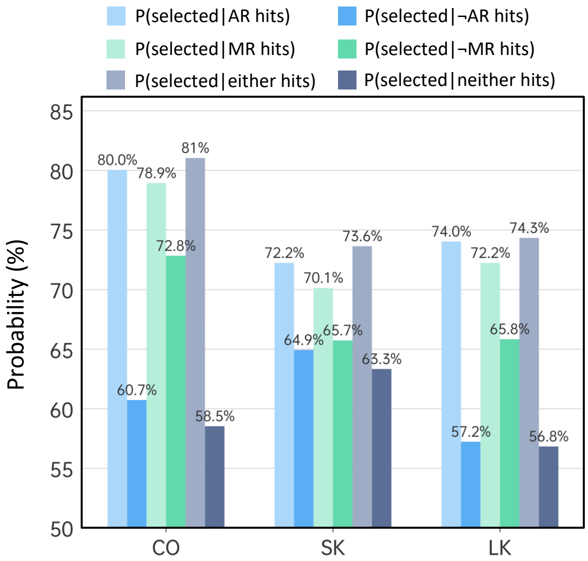

Figure 5: Conditional probabilities of the three answer candidates being selected conditioned on the exact occurrence of the answer in the rationales. There are six different conditions, including the answer candidate occurring in the automatic rationale (AR), the mechanistic rationale (MR), either rationale, and their respective negations. The abbreviations of strategy names are: CO = caption-only, SK = short-form knowledge, and LK = long-form knowledge.

strategy, we first extract all questions that this strategy can successfully answer by itself. Out of these questions, we show the success rate of the other two strategies. The results indicate that out of the questions answerable by one strategy, 15-19% cannot be answered by another strategy. Hence, each QA strategy is indispensable for the success of DIETCOKE.

The Rationalization Strategies. We incrementally integrate automatic and mechanistic rationales into DIETCOKE. The results are shown in Tab. 3. The additional baseline (#1) selects answer candidate randomly without LLM-based answer fusion. The ablation #2 selects answers using an LLM without rationales. #5 replaces the mechanistic rationale with a random sentence from the decision context.

Comparing #1 with #2, we observe minor improvements (0.6% and 1.1%) from the simplistic answer ensemble without rationales. #3 and #4 improve over #2 by 0.5% and 0.7% on OKVQA and by 0.3% and 1.1% on A-OKVQA, demonstrating both types of rationales are effective by themselves. Adding a random mechanistic rationale to #4, we obtain ablation #5, which retains the same performance on OKVQA but decreases by 0.6% on A-OKVQA. This highlights the importance of selecting proper mechanistic rationales, as wrong rationales are harmful. Most interestingly, leveraging both rationales, #6 attains the best performance, +1.2% / +1.4% over #2. Notably, the improvements are exactly the sum of the improvements attained by the individual rationales (#3 and #4), suggesting the two rationales are perfectly complementary and the ensemble over rationales is effective.

Finally, with Fig. 5, we analyze the relation between the rationale surface form containing exactly the answer candidate, which we refer to as a "hit", and the candidate being selected by the ensemble phase LLM. Similar to the finding of Tiong et al. (2022), we observe a positive correlation between the two events. This suggests that the surface form of the rationales plays a role in the decisions of the answer-selecting LLM. The largest differential is with caption-only, where having the answer in either rationale increases the selection probability by 22.5%[2]. In contrast, having the candidate in the rationales under SK or LK is not as important (probability differences of 17.5-10.3%). A possible reason is that the LLM conducts more complex operations on questions requiring external knowledge than simple surface-form matching.

4.4 Case Study

We show some example K-VQA questions in Fig. 6. To conserve space, we do not repeat the corresponding captions and knowledge details, available in Fig. 1. In strategy ensemble, we provide answer candidates with corresponding rationales and caption-generated QA pairs in the prompt. More examples can be found in Fig. 7 of the Appendix B.

In example (a), since "Brick walkway" appears in the captions, the answer from caption-only strategy is "Brick". However, as concrete is a more common street material, the answers from short-form and long-form knowledge are both "Concrete". Still, the model selects "Brick" as the final answer, possibly due to the strength of the rationale. In example (b), the answer from the caption-only strategy is "Muffin", but the its rationales are not proper explanations. The answer from the short-form knowledge strategy is correct and has sensible rationales, and it is selected as the final answer. In example (c), desert and savanna are both possible answers, which are hard to distinguish without

[2] P(selected | either hits)−P(selected | neither hits)=22.5%

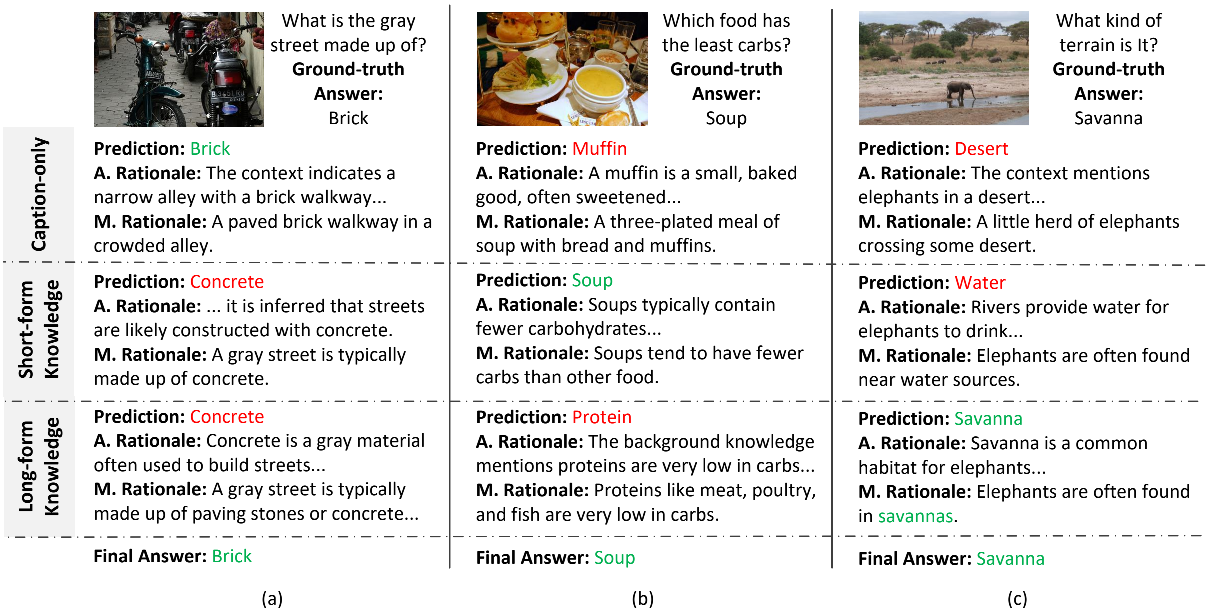

Figure 6: Examples from A-OKVQA val split where our method can help model select the correct answer. From top to bottom are the caption-only strategy, the short-form knowledge strategy, and the long-form knowledge strategy. To save space, please refer to Fig. 1 for the relevant captions and knowledge information.

visual information. The model may have been informed by the rationales from the second strategy mentioning rivers, which are more common in a savanna than in a desert.

5 Conclusion

We propose DIETCOKE, which ensembles several QA strategies for knowledge-based VQA. DIETCOKE first generates three types of decision contexts: image captions, short-form knowledge and long-form knowledge, and answers the question from each decision context separately. After that, DIETCOKE generates two types of rationales for each answer. An ensemble LLM selects the best answer from the rationales. DIETCOKE achieves state-of-the-art results on OK-VQA among comparable methods and showcases LLM-informed, rationale-based ensemble as an effective VQA method. DIETCOKE allows potential trade-off between compute and answer quality and may contribute to research on scaling LLMs with inference-time compute, a research direction pioneered by OpenAI-o1 (OpenAI, 2024).

Acknowledgments

We gratefully acknowledge the support by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

Limitations

In this paper, we generate three different types of decision contexts for each question, including captions, short-form knowledge, and long-form knowledge, where both types of knowledge are generated based on captions. However, in the process of converting images to captions, some visual details are inevitably lost, resulting in inaccurate captions. Therefore, the knowledge generated based on these captions may also be inaccurate. These inaccurate decision contexts may lead to incorrect answers. In extreme cases, if all answer candidates are wrong, our method will fail. Hence, improving the quality of decision contexts is an important direction for future research. Moreover, our method requires multiple calls to the LLM for inference, which allows inference-time scaling up of model capabilities but may take too long. Reducing the running time could be a direction for future research.

Although our work has achieved good results, it also inherits any existing biases of LLMs and their training datasets. Future work can focus on addressing these issues.

References

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

Leo Breiman. 2001. Random forests. In *Machine Learning*, pages 5–32.

Rui Cao and Jing Jiang. 2024. Knowledge generation for zero-shot knowledge-based vqa. *arXiv preprint arXiv:2402.02541*.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2024. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36.

Robert T. Clemen and Robert L. Winkler. 1985. Limits for the precision and value of information from dependent sources. *Operations Research*, 33(2):427–442.

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

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Jingjing Jiang and Nanning Zheng. 2023. Mixphm: Redundancy-aware parameter-efficient tuning for low-resource visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24203–24213.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. 2024. Exploring question decomposition for zero-shot vqa. *Advances in Neural Information Processing Systems*, 36.

Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4389–4400.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. *arXiv Preprint 2402.12875*.

Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022. Modular and parameter-efficient multimodal fusion with prompting. *arXiv preprint arXiv:2203.08055*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2022. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. 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*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31.

Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 451–468.

OpenAI. 2024. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2024-10-04.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *The Twelfth International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. 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*, pages 14974–14983.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge vqa. *arXiv preprint arXiv:2305.06407*.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023. Filling the image information gap for vqa: Prompting large language models to proactively ask questions. *arXiv preprint arXiv:2311.11598*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2023. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. *arXiv preprint arXiv:2310.13570*.

Xiaoying Xing, Mingfu Liang, and Ying Wu. 2023. Toa: Task-oriented active vqa. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

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

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.

A Experimental Details

We use the instruction-tuned LLM in generation of the knowledge text and the automatic rationale, and the LLM before instruction tuning in generation of the answer and the mechanistic rationale. For the standard version of the LLM, we use a greedy decoding strategy. For the instruction-tuned version of the LLM, we use top-k sampling, where k is set to 50. We set the temperature, the length penalty and the repetition penalty all to 1.0, and set the diversity penalty to 0.

B Prompt Templates

In this section, we demonstrate prompt templates for generating answer candidate, background knowledge, automatic rationale, and question-answering strategy fusion, as shown in Tab. 4, Tab. 5, Tab. 6 and Tab. 7, respectively. We use the same prompt templates across different datasets and models.

C Interactions Between Decision Contexts

To explore the effects of interactions between decision contexts, we conduct additional experiments using Mistral-7B by concatenating different decision contexts in one prompt. The result is 41.7% (44.7% resp.) on A-OKVQA val split (OKVQA test split resp.), which is much lower than 47.5% (49.2% resp.) of our method. Consistent with our observation, Cao and Jiang (2024) also concatenate captions and long-form knowledge in KGenVQA, and the result is 39.1% (45.4% resp.). We hypothesize that concatenating different decision contexts into a single prompt can introduce contradictory or redundant information, making it difficult for the model to effective utilize the information in the context to answer questions.

D Ensemble of Different Caption-only Strategies

We conduct experiments using three different image captioning models, BLIP (Li et al., 2022), InstructBLIP (Dai et al., 2023) and OFA (Wang et al., 2022) to generate different captions. We replace the short-form and long-form knowledge in our method with the captions generated by InstructBLIP and OFA. The result with Mistral-7B on A-OKVQA val split (OKVQA test split resp.) is 44.5% (46.7% resp.), much lower than 47.5% (49.2% resp.) from our method.

E Examples

In this section, we display more examples from OK-VQA where DIETCOKE can help model select the correct answer, as shown in Fig 7.

Understanding why the LLM selects certain answers is challenging for two main reasons. First, as recent research (West et al., 2023) indicates, LLMs exhibit behaviors drastically different from humans and may not fully comprehend the outputs they generate. That is, even if the LLM generates the right answer, it may not be able to explain why it selects a particular answer, or do other things that we expect a human who make the right choice would be able to do. Second, without ground-truth annotations for errors in the rationales, it is difficult to quantitatively assess if LLMs correctly detect illogical rationales or factual errors.

Despite these challenges, in Fig. 5, we analyze the relation between the rationale surface form containing exactly the answer candidate and the candidate being selected by the answer-selecting ensemble LLM. The positive correlation between the two events suggests that the surface form of the rationales plays a role in the decision of the answer-selecting LLM.

In addition, we provide three examples by asking the LLM to explain its choice to illustrate how the LLM selects answers. We use "Can you explain briefly how you select the final answer based on the three answer candidates and corresponding rationales?" as the instruction, and providing the final answer and three answer candidates with their rationales in the prompt. For question (a) in Fig. 7, the output is "1847 is the final answer because it marks the historical introduction of donuts to North America, supported by corresponding rationale." We speculate that the rationale behind the answer "1847" is more relevant to the question regarding the "first introduced" date, so the model selects "1847" as the final answer.

For question (b), the output is "The final answer 'Yacht' was selected because multiple rationales consistently describe a large luxury vessel in the scene, and the term 'yacht' aligns more specifically with that description than the more general term 'boat'." For question (c), the output is "I selected the final answer '8' because it provides the most precise and direct explanation by identifying the number of sides (8) of an octagonal stop sign."

We also provide some examples containing wrong rationales in Fig. 8.

Prompt Template for Generating Answer Candidate
Please answer questions according to the given context.
Context: *<Captions>/<Short-form Knowledge>/<Long-form Knowledge>*
<Question-Answer Pairs>
Question: *<Question>*
Answer:

Table 4: Prompt template for generating answer candidate. Choose one from three decision contexts.

Prompt Template for Generating Background Knowledge
User: You are going to answer questions according to the context: *<Captions>*
Assistant: Ok, please go ahead and ask your questions.
<Question-Answer Pairs>
User: *<Question>*
Assistant: I don't have enough knowledge to answer this question.
User: Please provide background knowledge related to this question (in a single sentence).

Table 5: Prompt template for generating background knowledge. We assume the model cannot answer the question using a caption-only strategy and ask it to generate relevant background knowledge.

Prompt Template for Generating Automatic Rationale
User: You are going to answer questions according to the context:
<Captions>/<Short-form Knowledge>/<Long-form Knowledge>
Assistant: Ok, please go ahead and ask your questions.
<Question-Answer Pairs>
User: *<Question>*
Assistant: *<Answer>*
User: Please explain the reasoning behind your answer in a single sentence.

Table 6: Prompt template for generating automatic rationale. Choose one from three decision contexts. We provide the answer candidate of corresponding strategy and ask the model to explain the reasoning behind it.

Prompt Template for Strategy Fusion
Please answer the question based on the most reasonable rationale.
Rationales:
1.$< A.\,Rationale\,of\,Answer_1 >$. $< M.\,Rationale\,of\,Answer_1 >$
2.$< A.\,Rationale\,of\,Answer_2 >$. $< M.\,Rationale\,of\,Answer_2 >$
3.$< A.\,Rationale\,of\,Answer_3 >$. $< M.\,Rationale\,of\,Answer_3 >$
$< Question - Answer\ Pairs >$
Question: $< Question >< Answer_1 >$ or $< Answer_2 >$ or $< Answer_3 >$?
Answer:

Table 7: Prompt template for ensemble. We first provide rationales for each answer, then ask the model to select a final answer from the answer candidates based on these rationales.

In what year was this dessert first introduced?
Ground Truth Answer: 1847

Captions: Variety of donuts on display on dessert table in restaurant. Many donuts are sitting on a tray in the window. A display case containing a close up of donuts. A variety of assorted pastries are on cakes. There are several different flavored donuts on the dessert plate...
Prediction: 1950
A. Rationale: The first introduction of this dessert can be traced back to the 1950s based on historical accounts of donuts gaining popularity in bakeries during that time.
M. Rationale: Variety of donuts on display on dessert table in restaurant.

Short-form Knowledge: Donuts are circular, sweet, fried pastries that have a hole in the middle and come in various flavors and toppings, with a history dating back to the 18th century.
Prediction: 1847
A. Rationale: Donuts, originating in Europe and brought to North America by Dutch settlers in 1847, are a circular, sweet, fried pastry with a hole in the middle, available in various flavors and toppings, and are displayed on tables in bakeries and shops.
M. Rationale: Donuts are circular, sweet, fried pastries that have a hole in the middle and come in various flavors and toppings, with a history dating back to the 18th century.

Long-form Knowledge: The first record of donuts being sold in the United States was in 1847 in New York City by a man named Hanson Crockett Gregory. In the United States, jelly-filled donuts gained popularity in the 1930s, when the first donut shops began to open.
Prediction: 1930
A. Rationale: The context mentions the jelly-filled donuts gained popular in the 1930s.
M. Rationale: In the United States, jelly-filled donuts gained popularity in the 1930s, when the first donut shops began to open.

Final Answer: 1847

(a)

What type of vessel is in the background?
Ground Truth Answer: Yacht

Captions: A white and gray yacht catamaran. Large yacht boat near large yacht in harbor. A boat in the river near a very large yacht. A white ship and a boat in the water. The motor yacht is in front of a yacht. A large yacht is a huge boat and one man. A big boat approaching a big brown & white yacht on the harbor...
Prediction: Boat
A. Rationale: Based on the context provided, it mentions a boat in the water multiple times. Therefore, the answer is "boat".
M. Rationale: A boat in the river near a very large yacht.

Short-form Knowledge: The context describes scenes of various boats, including yachts, anchored or sailing on the water, with some scenes taking place in the harbor or in front of buildings.
Prediction: Yacht
A. Rationale: The context describes scenes of yachts, anchored or sailing on the water.
M. Rationale: The context describes scenes of various boats, including yachts, anchored or sailing on the water, with some scenes taking place in the harbor or in front of buildings.

Long-form Knowledge:
A large yacht, which is a type of luxury watercraft, is mentioned to be present. A yacht is a luxury boat or ship used primarily for recreation, leisure, and pleasure cruising. Yachts come in various sizes and designs, ranging from small sailing vessels to large, motor-powered yachts.
Prediction: Yacht
A. Rationale: The context indicates a large yacht in the scene. So the vessel in the background is a yacht.
M. Rationale: A large yacht, which is a type of luxury watercraft, is mentioned to be present.

Final Answer: Yacht

(b)

How many sides does the sign have?
Ground Truth Answer: 8

Captions: A stop signage is sitting in front of a set of barricade barriers. A street stop sign in black in black and white. Black and white photograph of a stop sign next to a light. Stop sign stop sign and stop sign with some arrows and stripes. A stop sign sitting underneath a stop sign at the metal docks...
Prediction: None
A. Rationale: The context does not provide enough information to answer this question.
M. Rationale: Black and white photograph of a stop sign next to a light.

Short-form Knowledge: A stop sign is an octagonal traffic sign used to instruct drivers and pedestrians to come to a full stop before proceeding.
Prediction: Octagonal
A. Rationale: A stop sign is typically an octagonal traffic sign.
M. Rationale: A stop sign is an octagonal traffic sign used to instruct drivers and pedestrians to come to a full stop before proceeding.

Long-form Knowledge: A stop sign is a traffic sign that comes in the shape of an octagon. The sign features the words "STOP" written in large white letters on a red background. The eight sides of the octagonal shape of the sign make it easily recognizable to drivers, and the warning for motorists to come to a complete stop before reaching the marked intersection. Thus, a stop sign has eight sides with the word "STOP" written on one face of the octagon, making it a total of eight sides for the shape.
Prediction: 8
A. Rationale: The context mentions an octagon, which has 8 sides when counting all sides of the octagonal shape.
M. Rationale: Thus, a stop sign has eight sides with the word "STOP" written on one face of the octagon, making it a total of eight sides for the shape.

Final Answer: 8

(c)

Figure 7: Examples from OKVQA where our method can help model select the correct answer. From top to bottom are the caption-only strategy, the short-form knowledge strategy, and the long-form knowledge strategy. Incorrect answers are marked with red and correct answers are in green.

What nationality is represented by this crusty bread?

Ground Truth Answer: French/Italian

Captions: A sandwich with meat, ketchup and a glass of wine. This is a table with a couple of bread loafs on top of it. A sandwich, meat in a sub, and a drink. A striped place mat with lots of different foods and drinks. The sandwich that is sub and tomatoes sauce...

Prediction: Italian

A. Rationale: Based on the given contexts, the term "a table that has a sandwich and lots of different foods on the bread" indicates that the table is set for an Italian meal, as Italian cuisine often features various food items on the same dish or table.

M. Rationale: This is a table with a couple of bread loafs on top of it.

Short-form Knowledge: The term "sub" in sub sandwich refers to a submarine-shaped sandwich that is typically filled with various meats, vegetables, and sauces, and is sometimes served with a side of dipping sauce.

Prediction: Italian

A. Rationale: The given background knowledge indicates that subs are Italian sandwiches, and the mention of "crusty bread" is suggesting that type of bread commonly associated with Italian food.

M. Rationale: The term "sub" in sub sandwich refers to a submarine-shaped sandwich that is typically filled with various meats, vegetables, and sauces, and is sometimes served with a side of dipping sauce.

Long-form Knowledge: The crusty bread you're asking about could potentially represent French or Italian cuisine. In French cuisine, the term "baguette" or "baguette francaise" refers to a long, thin loaf of French bread with a crispy crust and soft interior. On the other hand, in Italian cuisine, the term "ciabatta" refers to a dense, chewy bread with a porous crumb and a crust that's crispy on the outside and soft on the inside.

Prediction: Italian

A. Rationale: Given the information provided, the term ciabatta refers to the described crusty bread in Italian cuisine, suggesting its Italian origin.

M. Rationale: The crusty bread you're asking about could potentially represent French or Italian cuisine.

Final Answer: Italian

(a)

What brand of soda is on the shelf?

Ground Truth Answer: Coca Cola

Captions: A bottle of soda is on the counter of a refrigerator. A kitchen with an open cabinet and a bottle of soda. A kitchen counter with a basket full of bottles and ingredients. A kitchen with a lot of bottles on the shelves. A collection of empty coke bottles in a run down kitchen...

Prediction: Coca Cola

A. Rationale: The context mentions a kitchen with soda bottles on glass shelves, and the brand "Coke" is specifically mentioned in another context.

M. Rationale: A collection of empty coke bottles in a run down kitchen.

Short-form Knowledge: A kitchen countertop is a surface where food preparation and serving activities take place, and it is often adorned with various containers and bottles, including those for drinks.

Prediction: Coca Cola

A. Rationale: I made an assumption based on the provided context that 'bottles' were referencing soda and 'brand' was specifically asking for Coca Cola.

M. Rationale: A kitchen countertop is a surface where food preparation and serving activities take place, and it is often adorned with various containers and bottles, including those for drinks.

Long-form Knowledge: In a kitchen setting, there are often various beverages, including sodas, kept on shelves or in a refrigerator. This context doesn't give any indication of the brand, only that "some bottles" are present on the glass shelves in a kitchen. To identify a brand, more information would be required, such as a product label or name mentioned in the context.

Prediction: None

A. Rationale: The context does not provide enough information to determine the specific brand of soda from the mentioned kitchen setting.

M. Rationale: This context doesn't give any indication of the brand, only that "some bottles" are present on the glass shelves in a kitchen.

Final Answer: Coca Cola

(b)

What type of soup will be made with this?

Ground Truth Answer: Vegetable

Captions: A healthy and small plate filled with many kinds of meat. A very bento bowl has rice and many vegetables. An app with rice and different foods inside it. A bowl of a variety of rice meal in a plastic yellow bowl. A close shot of a bowl of rice with meat. A plastic container with rice and vegetables arranged...

Prediction: Stew

A. Rationale: The given context describes a small plastic bowl filled with a stew pork meal, implying that the soup inside is a pork stew.

M. Rationale: A very bento bowl has rice and many vegetables.

Short-form Knowledge: A bento box is a traditional Japanese meal, usually served in a wooden box or a stacked set of containers, containing rice, fish or meat, and various side dishes.

Prediction: Soup

A. Rationale: The question provided no information about the ingredients in the bowl, but the context mentions that it is a bowl filled with different types of food, making it a suitable container for making soup.

M. Rationale: A bento box is a traditional Japanese meal, usually served in a wooden box or a stacked set of containers, containing rice, fish or meat, and various side dishes.

Long-form Knowledge: In general, a soup is a liquid dish that is made by simmering ingredients in a cooking pot. These ingredients can include vegetables, meat, grains, or legumes, and the liquid used to cook the soup can be water, stock, or milk. Common types of soup include vegetable soup, meatball soup, tomato soup, and rice noodle soup, among others.

Prediction: Soup

A. Rationale: Based on the given context, a soup is a liquid dish that can be made by simmering various ingredients such as vegetables, meat, grains, or legumes, and the question mentions nothing specific about a solid food served in a bowl, so it can be reasonably assumed that the question refers to a soup.

M. Rationale: In general, a soup is a liquid dish that is made by simmering ingredients in a cooking pot.

Final Answer: Soup

(c)

Figure 8: Examples from OKVQA where certain rationales are wrong. The wrong rationales and answers are marked in red, and correct answers are in green.