Boosting Knowledge-based Visual Question Answering with Structured Context Reasoning

Qiyou Liu[1], Yong Zhang[2], Jianjie Luo[1,*], Zhenguo Yang[1], Yi Yu[3]
[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[3]Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan
globarstarliu@gmail.com, yongzhang@szu.edu.cn, {jianjieluo, yzg}@gdut.edu.cn, yiyu@hiroshima-u.ac.jp

***Abstract*—Knowledge-based Visual Question Answering aims to answer questions about an image by integrating external knowledge with visual and textual information. Recent approaches often rely on in-context learning to prompt Large Language Models (LLMs) with multimodal context in a zero-shot or few-shot manner. However, we observe that directly concatenating heterogeneous visual descriptions and retrieved knowledge into long, unstructured prompts often degrades reasoning performance, due to both excessive irrelevant context and the lack of explicit relational structure. In this paper, we propose an LLM-based Structured Context Reasoning (SCoRe) framework that infers both explicit and implicit relationships for prediction. SCoRe consists of three stages: Context Acquisition, which generates diverse visual notes and retrieves explicit knowledge via an efficient two-stage multimodal retrieval strategy; Context Selection, which filters relevant visual, explicit, and implicit knowledge using LLM-guided selection; and Context Compression, which performs Relational Logic Distillation (RLD) to transform raw text into explicit entity-relation triplets. These relational triplets serve as a concise and structured prompt for final answer prediction. Extensive experiments on the OK-VQA and A-OKVQA benchmarks demonstrate that SCoRe consistently outperforms state-of-the-art methods.**

***Index Terms*—Knowledge-based Visual Question Answering, Structured Context Reasoning.**

I. Introduction

Knowledge-based Visual Question Answering (KB-VQA) answers questions by integrating external knowledge with visual and textual information. Compared with conventional VQA, KB-VQA often involves implicit concepts and multi-step reasoning, making effective organization of heterogeneous information crucial for accurate prediction.

Recent works [1]–[4] increasingly leverage Large Language Models (LLMs) to address KB-VQA in zero-shot or few-shot settings by concatenating image descriptions, retrieved knowledge, and task instructions into a single prompt, as illustrated in Fig. 1 (a). Although being effective in some cases, this unstructured prompting paradigm presents inherent limitations. First, the prompt often contains redundant or weakly relevant information, which distracts the model from key evidence. Second, unstructured context fails to explicitly represent relationships among entities, forcing LLMs to implicitly infer complex logical dependencies during reasoning. These issues are particularly pronounced in KB-VQA, where answering questions often depends on understanding relations rather than

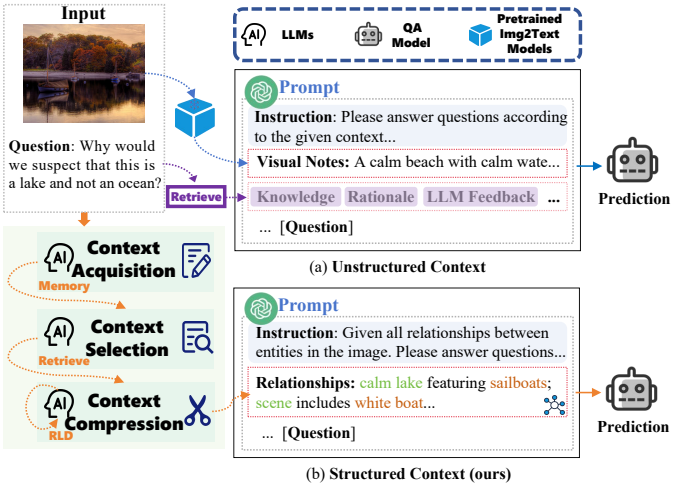

Fig. 1. Two paradigms for LLM-based KB-VQA: (a) Unstructured context reliance, where raw evidence is concatenated into flat, redundant prompts; (b) Our structured context reasoning, which distills multimodal inputs into relational triplets through progressive acquisition, selection, and compression.

isolated facts. Several prior methods attempt to alleviate these problems through knowledge selection [5], rationale extraction [6], or heuristic filtering [7]. However, such approaches still operate over a flat textual context and do not explicitly model relational structure.

To this end, we propose Structured Context Reasoning (SCoRe), a framework that explicitly transforms heterogeneous multimodal inputs into compact relational representations tailored for LLM-based reasoning, as illustrated in Fig. 1 (b). Instead of treating context as a flat sequence of text, SCoRe progressively constructs a structured reasoning substrate through three stages. First, the Context Acquisition stage generates diverse visual notes and retrieves explicit external knowledge using an efficient two-stage multimodal retrieval strategy. Second, the Context Selection stage leverages LLMs to identify and retain only information that is relevant to the question, while also eliciting implicit knowledge aligned with the visual content. Finally, the Context Compression stage implements Relational Logic Distillation (RLD) guided through a systematic logical decomposition process, which distills flat text into structured entity-relation triplets, yielding a compact and logic-oriented context for answer prediction. Experiments on OK-VQA and A-OKVQA demonstrate that SCoRe consistently outperforms state-of-the-art zero-shot and

*J. Luo is the corresponding author.

few-shot methods. The results indicate that explicitly structuring context into relational representations improves reasoning effectiveness beyond simply increasing or filtering textual context. We show that context structure, rather than context quantity, is the key factor for effective KB-VQA.

In summary, our contributions are as follows:

- We propose SCoRe, a structured context reasoning framework for KB-VQA that reformulates heterogeneous multimodal information into compact relational representations, enabling more effective LLM-based reasoning.
- We design a two-stage progressive multimodal retrieval strategy that significantly reduces the retrieval search space while preserving high recall of relevant knowledge.
- We propose RLD, a logic-driven inference mechanism that models entity relations to transform evidence into a structured, reasoning-oriented context.
- We conduct extensive experiments on OK-VQA and A-OKVQA, achieving consistent improvements over strong baselines and validating the effectiveness of structured context representations for KB-VQA.

II. Related Work

Knowledge-based VQA. Knowledge-based VQA integrates external knowledge with multimodal inputs to answer questions. Traditional methods [8], [9] relied on supervised learning with retrieved knowledge (e.g., Wikipedia), but faced limitations in retrieval accuracy and generalization. With the rise of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs), quite a few methods [1], [7] have started to use LLMs or LVLMs as reasoning models, leveraging the knowledge they contain to infer answers. The basic method [2] directly used image descriptions to prompt LLMs. For questions that required comprehensive reasoning, the LLMs may struggle to arrive at answers based solely on image descriptions. The subsequent method KGenVQA [5] leveraged LLMs to generate relevant knowledge, then combined the image descriptions to answer the question. This may lead to the introduction of a large amount of redundant information. The method DIETCOKE [6] leveraged the LLMs to extract rational sentences from the context. This helped LLMs focus more effectively on key information, but improper extraction may lead to the loss of some important details. These methods based on LLMs all employed unstructured context, which resulted in incoherent logic. Moreover, methods such as DIETCOKE [6] relied solely on the LLMs' knowledge for question answering, whereas we retrieve explicit knowledge following the PREFLMR [10] and combine it with LLMs' knowledge to predict the answer.

Relation Extraction. Relation Extraction (RE) identifies structured entity dependencies to provide a logical foundation for reasoning. MUMRC [11] recasts multimodal triplet extraction as machine reading comprehension, while TEVLA [12] utilizes generative text augmentation for robust vision-language alignment. Recent LLM-based methods like QA4RE [13] elicit zero-shot extraction by aligning RE with common instruction tasks, and OSLLM [14] improves consistency via a retrieve-reason-refine framework. Additionally, ERA-Cot [15] underscores that analyzing entity relationships is vital for enhancing logical depth. Unlike methods focused on standalone extraction accuracy, our SCoRe framework implements Relational Logic Distillation as a systematic logical decomposition process. It distills heterogeneous evidence into structured relational triplets, effectively alleviating cognitive overload and facilitating final answer prediction in KB-VQA.

III. Methodology

In this section, we introduce SCoRe in detail. SCoRe includes three stages: Context Acquisition (CA), Context Selection (CS), and Context Compression (CC). Specifically, the CA stage is responsible for retrieving explicit knowledge and converting image information into multi-aspect visual notes. The CS stage retrieves implicit knowledge and extracts the relevant information to obtain a concise context. Subsequently, the CC stage combines such information to extract structured relations. Finally, the QA model produces answers using these relations. The overview of SCoRe is shown in Fig. 2.

A. Context Acquisition

The Context Acquisition stage aims to construct a comprehensive but decomposed evidence pool from both visual inputs and external knowledge sources.

Visual Evidence. To preserve complementary visual cues, we represent the image using three parallel views: object-level attributes, image captions, and OCR-extracted text. Object attributes are extracted using a pre-trained detector, captions are generated via question-guided captioning, and OCR strings are obtained from the image text. Together, these components form a set of visual notes that explicitly separate different types of visual evidence.

Explicit Knowledge Retrieval. Consistent with previous work, we employ the Google Search corpus and Wikipedia as explicit knowledge bases for OK-VQA and A-OKVQA, respectively. To address the retrieval efficiency bottleneck caused by the massive scale of the Wikipedia, we design a two-stage progressive multimodal retrieval strategy, as illustrated in Fig. 3. Specifically, we first utilize a text encoder F_L and a visual encoder F_V to encode the question q and the image i, respectively. After that, through a mapping layer F_M, the original visual features are projected into the text feature space to obtain aligned visual features. The text features and the aligned visual features are concatenated along their dimensions to form a cross-modal query vector Q:

$$Q = \text{Concat}(F_L(q), F_M(F_V(i))) \in \mathbb{R}^{(l_q+l_i)\times d_L}, \quad (1)$$

where l_q denotes the sequence length of question q, and l_i denotes the sequence length of the mapped visual features.

First Stage: Title-level multimodal coarse retrieval. Using the same text encoder F_L as used for question encoding, we encode title t ($t \in Title$) to generate a title feature vector:

$$T = F_L(t) \in \mathbb{R}^{l_t \times d_L}, \quad (2)$$

where l_t denotes the sequence length of title t. Based on the late interaction mechanism, finegrained relevance scores

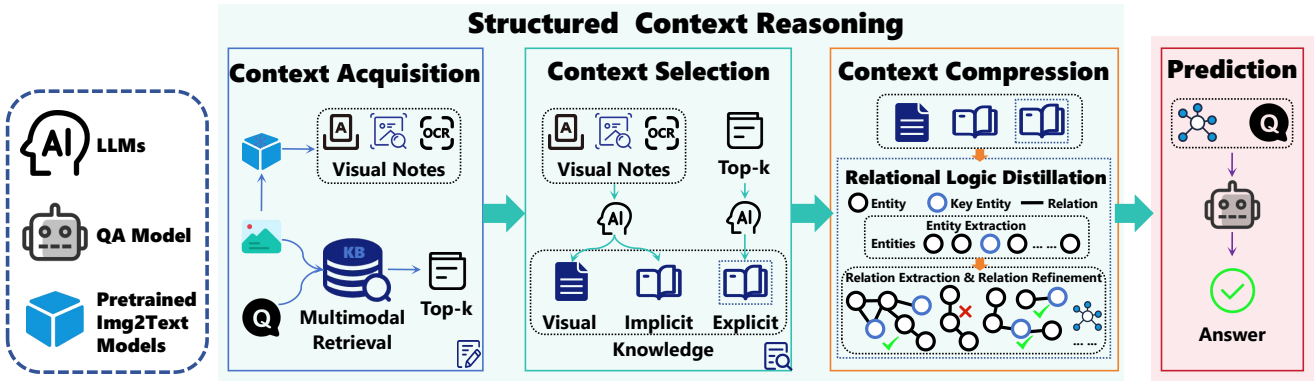

Fig. 2. The proposed LLM-based Structured Context Reasoning (SCoRe) method.

between the query vector Q and each title t are computed as:

$$S(\overline{q},t) = S((q,i),t) = \sum_{x=1}^{l_Q} \max_{y=1}^{l_T} Q_x T_y^{\mathsf{T}}, \tag{3}$$

where $l_Q = l_q + l_i$ denotes the sequence length of the query vectors, Q_x denotes the x-th query embedding vector, and T_y denotes the y-th embedding vector. The relevance scores for all titles are sorted in descending order, and the top k_1 titles corresponding to the highest scores are selected to form the candidate title set:

$$T_{\text{top-}k_1} = \arg \max_{\substack{T' \subseteq Title \\ |T'|=k_1}} \{S(\overline{q},t) \mid t \in Title\}. \tag{4}$$

Finally, we extract the union of all candidate passage subsets $P_{candidate}$ corresponding to the M samples:

$$P_{candidate} = \bigcup_{m=1}^{M} \left\{ \bigcup_{t \in T_{\text{top-}k_1}} P_t \right\}. \tag{5}$$

At this point, the size of $P_{candidate}$ is much smaller than that of the full passage set $Passage$, yielding an efficient compression of the retrieval scope.

Second stage: Paragraph-level multimodal fine-grained retrieval. We take $P_{candidate}$ as the retrieval target, reuse the same relevance metric, and achieve precise recall of the target knowledge paragraphs. Specifically, we encode p ($p \in P_{candidate}$) using text encoder F_L:

$$P = F_L(p) \in \mathbb{R}^{l_p \times d_L}, \tag{6}$$

where l_p denotes the sequence length of p. Applying the relevance scoring function defined in the first stage, the relevance score between the query vector Q and paragraph feature P is computed as:

$$S(\overline{q},p) = S((q,i),p) = \sum_{x=1}^{l_Q} \max_{y=1}^{l_P} Q_x P_y^{\mathsf{T}}, \tag{7}$$

where P_y denotes the y-th embedding vector. Finally, the relevance scores of the candidate paragraphs are sorted in descending order, and the knowledge paragraphs corresponding to the top-k_2 scores are selected to form the target knowledge fragment set for a single sample as:

$$P_{\text{top-}k_2} = \arg \max_{\substack{P' \subseteq P_{candidate} \\ |P'|=k_2}} \{S(\overline{q},p) \mid p \in P_{candidate}\}, \tag{8}$$

where $P_{\text{top-}k_2}$ represents the target knowledge paragraph set.

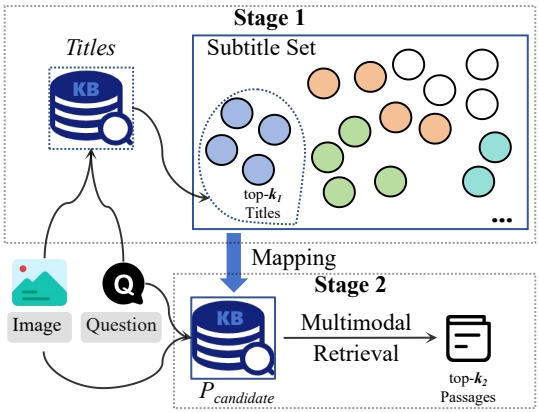

Fig. 3. The two-stage progressive multimodal retrieval strategy.

B. Context Selection

The Context Selection stage focuses on reducing evidence redundancy while preserving question-relevant semantics. LLMs are utilized to select visual knowledge and explicit knowledge from visual notes and top-k_2 explicit knowledge passages, respectively.

Generally, for an image I_i corresponding to question Q_i, along with the caption C_i, objects O_i, OCR string S_i, and explicit knowledge passages $P^i_{\text{top-}k_2}$ obtained from the CA stage, the selected visual knowledge is $VK_i = \mathcal{LLM}[\mathcal{P}(C_i, O_i, S_i, Q_i)]$, the selected explicit knowledge is $EK_i = \mathcal{LLM}[\mathcal{P}(C_i, Q_i, P^i_{\text{top-}k_2})]$, where $\mathcal{P}$ denotes the prompt template for this scene. Specifically, we feed both image captions and the question to retrieve implicit knowledge from LLMs to ensure alignment: $IK_i = \mathcal{LLM}[\mathcal{P}(C_i, Q_i)]$.

C. Context Compression

To bridge the gap between unstructured evidence and logical reasoning, we propose Relational Logic Distillation (RLD) in the Context Compression stage. The RLD is operationalized through a systematic logical decomposition procedure, which guides the LLM to transform textual evidence into explicit relational structures. RLD consists of three key steps: entity extraction, relation extraction, and relation refinement.

In the first step, we initialize the entity set E_{Vi} using the detected objects O_i:

$$E_{Vi} = \{e^1_{Vi}, ..., e^k_{Vi}\} \leftrightarrow O_i. \tag{9}$$

To avoid missing critical entities, we then use an LLM to separately extract the explicit and implicit knowledge entities from the sentence as follows: $E_{Ei} = \mathcal{LLM}[\mathcal{P}(VK_i, EK_i)], E_{Ii} = \mathcal{LLM}[\mathcal{P}(IK_i)]$. In consequence, the final explicit entity list can be expressed as $E_i^e = E_{Vi} \cup E_{Ei}$, the implicit entity list can be expressed as $E_i^i = E_{Vi} \cup E_{Ii}$.

In the second step, we extract the explicit and implicit relationships between entity pair (e_p, e_q) where $e_p \neq e_q$. The relation is denoted as triplets (e_p, r, e_q) where r is the relation of entities e_p and e_q. Specifically, for the i-th sample, its corresponding entity set E_i^e and E_i^i, the initial explicit relation set $ER_i^{init} = \{..., (e_p, r_E, e_q), ...\}$ (e_p, $e_q \in E_i^e$ and $e_p \neq e_q$) and the implicit relation set $IR_i^{init} = \{..., (e_p, r_I, e_q), ...\}$ (e_p, $e_q \in E_i^i$ and $e_p \neq e_q$) are formulated as:

$$ER_i^{init} = \mathcal{LLM}[\mathcal{P}((VK_i, EK_i), E_i^e)], \tag{10}$$

$$IR_i^{init} = \mathcal{LLM}[\mathcal{P}(IK_i, E_i^i)]. \tag{11}$$

Therefore, the initial relations $R_i^{init} = ER_i^{init} \cup IR_i^{init}$.

In the final relation refinement step, we utilize an LLM to extract a list of key entities from the image caption and the question, and then filter the initial relation set by removing irrelevant relational triples. Specifically, for the visual knowledge VK_i corresponding to question Q_i, the key entity list can be denoted as: $KE_i = \mathcal{LLM}[\mathcal{P}(VK_i, Q_i)]$. The final relations can be expressed as:

$$\begin{aligned} R_i = \big\{ (e_p, r, e_q) \;\big|\; (e_p, r, e_q) \in R_i^{init} \wedge [e_p, e_q \in KE_i \\ \vee\, (\exists e_k \in KE_i, (e_p \overset{*}{\leftrightarrow} e_k) \;\vee\; (e_q \overset{*}{\leftrightarrow} e_k))] \big\}, \end{aligned} \tag{12}$$

where $\overset{*}{\leftrightarrow}$ denotes that the two entities can reach each other through one or more relations.

D. Prediction

In the final phase, the QA model predicts the answer by conditioning on the structured relation set R_i and the question Q_i. By replacing unstructured context with explicit relational representations, the model can focus on reasoning over entity dependencies rather than on organizing raw information.

IV. Experiments

A. Experiment Settings

Datasets. We conduct experiments on the two popular KB-VQA datasets: Outside Knowledge VQA (OK-VQA) [25] and Augmented OK-VQA (A-OKVQA) [26]. Following EF-VQA [4], experimental evaluations are performed on the test set of OK-VQA, and the validation set of A-OKVQA.

Implementation Details. DeepSeek-V3 is used as the context processor, while Mistral-7B and Gemma-7B are employed as the QA models. For DeepSeek-V3, we set the temperature to 0.7 and the presence penalty to 0.2. For Mistral-7B and Gemma-7B, we adopt a greedy decoding strategy. DeepSeek-V3 is accessed via API, whereas Mistral-7B and Gemma-7B are locally deployed on two RTX 3090 24 GB GPUs. We use VinVL [27] for object detection, apply the captioning method from PNP-VQA [16] to generate question-guided captions, and extract image text with the Google OCR API. In addition, we adopt the pre-trained PREFLMR [10] as the backbone for multimodal retrieval.

TABLE I
COMPARISON WITH SOTA METHODS ON THE OK-VQA AND A-OKVQA DATASETS. THE BEST SCORES ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST SCORES ARE HIGHLIGHTED IN UNDERLINE.

Method	QA Model	Shot Number	OK-VQA	A-OKVQA
LLM-based Zero-shot Methods				
PICa [1]	GPT-3_{175B}	0	17.1	-
PNP-VQA [16]	UnifiedQA$_{11B}$	0	35.9	-
Img2LLM [2]	OPT$_{175B}$	0	45.6	42.9
LAMOC [17]	FLAN-T5-XXL$_{11B}$	0	40.3	37.9
RQ prompt [18]	GPT-3_{175B}	0	46.4	43.2
Emu [19]	Llama$_{13B}$	0	38.2	-
ZVQAF [20]	FLAN-T5$_{11B}$	0	40.5	37.1
KGenVQA [5]	UnifiedQA$_{11B}$	0	45.4	39.1
EF-VQA [4]	GPT-3.5$_{175B}$	0	43.7	42.1
L2A [21]	GPT-3.5-turbo	0	46.2	<u>48.5</u>
PLLMKI [22]	Llama$_{7B}$	0	28.4	24.8
DIETCOKE [6]	Gemma$_{7B}$	0	47.6	47.3
DIETCOKE [6]	Mistral$_{7B}$	0	<u>49.2</u>	47.5
SCoRe (ours)	Gemma$_{7B}$	0	49.4	50.3
SCoRe (ours)	Mistral$_{7B}$	0	**52.1**	**52.4**
LLM-based Few-shot Methods				
PICa [1]	GPT-3_{175B}	1	40.8	-
PICa [1]	GPT-3_{175B}	4	45.4	-
PICa [1]	GPT-3_{175B}	16	48.0	-
MM-R [23]	GPT-4-32k	20	40.0	-
VCTP [3]	OPT$_{66B}$	8	44.6	46.4
PCPA [24]	Llama$_{7B}$	16	53.6	52.3
PLLMKI [22]	Llama$_{7B}$	8	<u>54.1</u>	<u>52.9</u>
EF-VQA [4]	GPT-3.5$_{175B}$	8	48.4	48.6
EF-VQA [4]	GPT-3.5$_{175B}$	16	51.2	50.1
SCoRe (ours)	Mistral$_{7B}$	4	53.2	53.0
SCoRe (ours)	Mistral$_{7B}$	8	54.8	53.6
SCoRe (ours)	Mistral$_{7B}$	16	**55.7**	**54.9**

Baselines. The methods we compare are categorized as follows: (1) LLM-based zero-shot methods: PNP-VQA [16], Img2LLM [2], LAMOC [17], RQprompt [18], Emu [19], ZVQAF [20], KGenVQA [5], L2A [21], and DIETCOKE [6]. (2) LLM-based few-shot methods: PICa [1], MM-R [23], VCTP [3], PCPA [24], PLLMKI [22], and EF-VQA [4]. We follow official evaluation protocols and report VQA scores for the datasets. The accuracy metric is defined as $VQA\ score = min(1, T(a)/3)$, where $T(a)$ is the number of times the predicted answer appears in groundtruth. Notably, we choose the direct answer evaluation of A-OKVQA dataset.

Zero-shot Settings. We follow Img2LLM [2] to generate QA pairs from image captions as demonstrations for zero-shot scenario. Throughout the entire process, the answer and any other samples remain untouched.

Few-shot Settings. We select the top-n training samples as exemplars based on the average cosine similarity between CLIP-based image and question embeddings for reasoning.

B. Performance of the Approaches

The experimental results summarized in Table I show that SCoRe consistently outperforms all baselines across both zero-shot and few-shot settings on OK-VQA and A-OKVQA. In the zero-shot scenario, SCoRe (Mistral-7B) achieves 52.1%

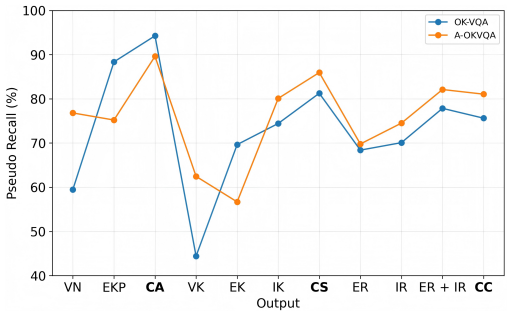

Fig. 4. The Pseudo Recall of intermediate outputs.

TABLE II
ABLATION STUDY RESULTS OF KEY MODULES ON OK-VQA DATASET.

VN	EKP	CS	CC	OK-VQA
✓				47.5
✓	✓			49.0
✓	✓	✓		51.2
✓	✓	✓	✓	**52.1**

and 52.4%, surpassing the previous best method, DIETCOKE, by a substantial margin. For the few-shot settings, SCoRe demonstrates steady performance gains as the shot count increases. At 16-shot, it reaches peak accuracies of 55.7% and 54.9%, establishing a new state-of-the-art. These consistent improvements across multiple benchmarks and settings underscore that transforming heterogeneous context into structured relational representations is more effective for knowledge-intensive reasoning than relying on unstructured text.

We also compare the time cost of SCoRe and DIETCOKE on Mistral-7B. SCoRe takes approximately 17.5 seconds per sample, whereas DIETCOKE requires 29.5 seconds. For visual question answering tasks, such time costs are acceptable while achieving improved accuracy.

C. Ablation Studies

Effectiveness of Proposed Modules. We conduct an ablation study on the OK-VQA dataset to validate the effectiveness of each module in SCoRe. As shown in Table II, we gradually incorporate each component to evaluate the performance improvement. Using only visual notes (VN) yields 47.5%. Adding explicit knowledge passages (EKP) further improves the result to 49.0%. With the introduction of the CS module, the score increases to 51.2%. Finally, equipping with the CC module achieves the best performance of 52.1%. These results demonstrate that each proposed module contributes to the final performance, and the full model with all components delivers the optimal reasoning ability.

Effects of Parameters. We use the Pseudo Recall (PR) to evaluate the impact of k_1 and k_2 on the retrieval results, where the PR is defined as the proportion of knowledge passages that contain the correct answer among all the retrieved passages. Experimental results on A-OKVQA validation set (1145 samples) are presented in Table III and Table IV. Table III shows that performance is already strong when k_1=1 and improves steadily as the number of titles increases. Beyond k_1=50 the gain tapers off, and PR@20 is almost identical for k_1=100 and k_1=200. Balancing efficiency and accuracy, we set k_1=200, k_2=10. Table IV lists the counts of candidate passages for different k_1 values. We observe that our method compresses the retrieval counts by tens to hundreds of times while still delivering promising results.

TABLE III
THE PSEUDO RECALL ACROSS DIFFERENT k_1 AND k_2.

top-k	k_1=1	k_1=10	k_1=20	k_1=50	k_1=100	k_1=200
k_2=5 (PR@5)	55.02	56.41	58.77	63.58	63.58	64.62
k_2=10 (PR@10)	66.72	68.64	68.90	73.71	74.14	75.19
k_2=20 (PR@20)	76.76	78.07	77.99	81.74	82.88	82.88

TABLE IV
THE NUMBER OF CANDIDATE PASSAGES FOR DIFFERENT k_1.

	ALL	k_1=1	k_1=10	k_1=20	k_1=50	k_1=100	k_1=200
Num	21,015,324	8,746	72,969	135,835	294,726	560,324	988,434

Information Loss of Intermediate Output. As SCoRe adopts a pipeline-based architecture, each stage of content generation or extraction inevitably alters and partially loses information. Fig. 4 shows the PR of intermediate outputs, including Visual Notes (VN), Explicit Knowledge Passages (EKP), CA (VN + EKP), Visual Knowledge (VK), Explicit Knowledge (EK), Implicit Knowledge (IK), CS (VK + EK + IK), Explicit Relations (ER), Implicit Relations (IR), ER + IR, and CC (refined ER + IR). As can be seen from the figure, the PR at the initial CA stage is approximately 10 percentage points higher than that at the final CC stage on two datasets. However, the actual performance of the CA is lower than the CC. This is because CA contains substantial redundant and noisy information, which impairs the model's judgment. In contrast, the structured information in the CC stage enables the model to better understand the relations between entities, thereby deriving accurate answers.

D. Case study

As shown in Fig. 5, we select some representative cases from OK-VQA to compare SCoRe with KGenVQA (a-b, zero-shot) and EF-VQA (c-d, few-shot). Baselines relying on unstructured implicit knowledge often suffer from noisy or incorrect reasoning. In contrast, SCoRe extracts structured relation triplets via context reasoning, integrating visual, explicit and implicit knowledge. SCoRe outputs accurate answers (e.g., "snow", "celebration") by capturing key info and logical connections, verifying its superiority in KB-VQA.

Despite its effectiveness, SCoRe occasionally encounters limitations, as illustrated in the failure cases. In Fig. 6 (a), although the explicit knowledge correctly retrieves "1892", the model outputs "1886" due to a conflict between explicit and implicit relations. The LLM's internal knowledge regarding the formula's invention (1886) overrides the specific brand founding date (1892) present in the retrieved context, leading to an incorrect reasoning priority. In Fig. 6 (b), the model provides a generic answer "birds" instead of specific species like "cardinal". This failure is attributed to insufficient finegrained

Fig. 5. The qualitative comparison of KGenVQA with zero-shot settings (a-b) and EF-VQA with few-shot settings (c-d).

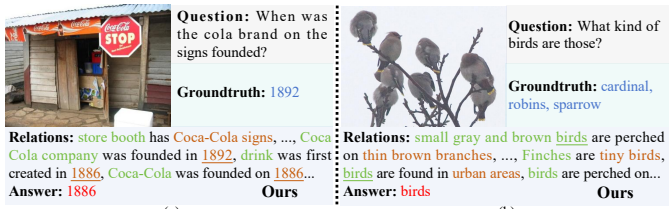

Fig. 6. Qualitative case study of SCoRe failure modes.

visual knowledge, while the SCoRe framework identifies the presence of birds and their basic colors, it lacks the specialized ornithological features (e.g., subtle beak shapes or plumage patterns) necessary for precise taxonomic classification.

V. Conclusion

In this paper, we propose an LLM-based Structured Context Reasoning (SCoRe) framework, which extracts structured relationships for knowledge-based visual question answering. SCoRe addresses two pivotal challenges by structured context reasoning to refine the original information: 1) Excessive inputs cognitively overload LLMs, and 2) Difficulty in reasoning over unstructured input. In addition, SCoRe introduces a two-stage progressive multimodal retrieval strategy that dramatically improves retrieval efficiency while maintaining high accuracy. Extensive experiments demonstrate that SCoRe outperforms state-of-the-art methods.

Acknowledgments

This work is supported by the Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010237), and a grant from CMA Communication and Outreach Centre (China Meteorological News Press) Meteorological Converged Media Technology Innovation and Application Open Laboratory (No. QXRM-ZD-2508).

References

[1] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *AAAI*, 2022.

[2] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *CVPR*, 2023.

[3] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, Z. Sun, D. Gutfreund, and C. Gan, "Visual chain-of-thought prompting for knowledge-based visual reasoning," in *AAAI*, 2024.

[4] P. Qiang, H. Tan, and X. Li, "Enhancing few-shot kb-vqa with panoramic image captions guided by large language models," *Neurocomputing*, 2025.

[5] R. Cao and J. Jiang, "Knowledge generation for zero-shot knowledge-based vqa," in *EACL*, 2024.

[6] M. Li, H. Li, Z. Du, and B. Li, "Diversify, rationalize, and combine: Ensembling multiple qa strategies for zero-shot knowledge-based vqa," in *EMNLP*, 2024.

[7] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *CVPR*, 2023.

[8] W. Lin and B. Byrne, "Retrieval augmented visual question answering with outside knowledge," in *EMNLP*, 2022.

[9] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based vqa," in *AAAI*, 2022.

[10] W. Lin, J. Mei, J. Chen, and B. Byrne, "PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers," in *ACL*, 2024.

[11] Q. Chen, D. Zhang, S. Li, and G. Zhou, "A unified mrc framework with multi-query for multi-modal relation triplets extraction," in *ICME*, 2023.

[12] J. Chen, Q. Guo, K. C. Cheung, M. Liang, and D. Chen, "Tevla: Text-oriented enhancement for vision-language alignment in relation extraction," in *ICME*, 2025.

[13] K. Zhang, B. Jimenez Gutierrez, and Y. Su, "Aligning instruction tasks unlocks large language models as zero-shot relation extractors," in *ACL*, 2023.

[14] J. Zhou, Y. Shan, M. Wu, F. Hu, L. Zheng, and X. Wang, "Osllm: A retrieve-reason-refine framework for multi-domain relation extraction with large language models," in *ICME*, 2025.

[15] Y. Liu, X. Peng, T. Du, J. Yin, W. Liu, and X. Zhang, "Era-cot: Improving chain-of-thought through entity relationship analysis," in *ACL*, 2024.

[16] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," in *EMNLP*, 2022.

[17] Y. Du, J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Zero-shot visual question answering with language model feedback," in *ACL*, 2023.

[18] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, "Improving zero-shot visual question answering via large language models with reasoning question prompts," in *ACM MM*, 2023.

[19] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang *et al.*, "Emu: Generative pretraining in multimodality," in *ICLR*, 2024.

[20] C. Liu, C. Wang, Y. Peng, and Z. Li, "Zvqaf: Zero-shot visual question answering with feedback from large language models," *Neurocomputing*, 2024.

[21] X. Xing, P. Xiong, L. Fan, Y. Li, and Y. Wu, "Learning to ask denotative and connotative questions for knowledge-based vqa," in *EMNLP*, 2024.

[22] Z. Hu, P. Yang, F. Liu, Y. Meng, and X. Liu, "Prompting large language models with knowledge-injection for knowledge-based visual question answering," *Big Data Mining and Analytics*, 2024.

[23] M. Khademi, Z. Yang, F. V. Frujeri, and C. Zhu, "Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering," in *EMNLP*, 2023.

[24] Z. Hu, P. Yang, Y. Jiang, and Z. Bai, "Prompting large language model with context and pre-answer for knowledge-based vqa," *Pattern Recognition*, 2024.

[25] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019.

[26] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," in *ECCV*, 2022.

[27] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang *et al.*, "Vinvl: Revisiting visual representations in vision-language models," in *CVPR*, 2021.