Knowledge Image Matters: Improving Knowledge-Based Visual Reasoning with Multi-Image Large Language Models

Guanghui Ye[1], **Huan Zhao**[1]*, **Zhixue Zhao**[2], **Yang Liu**[1], **Xupeng Zha**[1], **Zhihua Jiang**[3]
[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Department of Computer Science, University of Sheffield, UK
[3]Department of Computer Science, Jinan University, China
{yghui,hzhao}@hnu.edu.cn

Abstract

We revisit knowledge-based visual reasoning (KB-VR) in light of modern advances in multimodal large language models (MLLMs), and make the following contributions: (i) We propose **V**isual **K**nowledge **C**ard (**VKC**) – a novel image that incorporates not only internal visual knowledge (e.g., scene-aware information) detected from the raw image, but also external world knowledge (e.g., attribute or object knowledge) produced by a knowledge generator; (ii) We present **VKC**-enhanced **M**ulti-**I**mage **R**easoning (**VKC-MIR**) – a four-stage pipeline which harnesses a state-of-the-art scene perception engine to construct an initial VKC (Stage-1), a powerful LLM to generate relevant domain knowledge (Stage-2), an excellent image editing toolkit to introduce generated knowledge into an iteratively-edited VKC (Stage-3), and finally, an emerging multi-image MLLM to solve the VKC-enhanced task (Stage-4). By performing experiments on three popular KB-VR benchmarks, our approach achieves new state-of-the-art results compared to previous top-performing models. Our code is available at: https://github.com/yyy1103/VKC.

1 Introduction

Knowledge-based visual reasoning (KB-VR) (Xu et al., 2024a; Song et al., 2024; Jin et al., 2023; Shao et al., 2023; Chen et al., 2024b) remains a challenging task, as it requires machines not only to understand the concepts and relationships of visual scenes, but also to associate them with external world knowledge to perform a chain of reasoning on open-world questions. As illustrated in Fig.1 (a) and (b), for KB-VR tasks composed of image (V), question (Q), and answer (A), image captioning (Liu et al., 2024c; Su and Gou, 2024) and scene graph generation (Zhang et al., 2024; Kim et al., 2024) are often considered foundational tasks to validate the model's general understanding of each image. For example, scene graphs provide fine-grained spatial information about an image. In Natural Language Processing (NLP), knowledge is generally represented as structured triples or graphs (Schwenk et al., 2022; Chen et al., 2023, 2024b), which facilitate representation learning. However, it is under-explored to depict knowledge in other modalities (e.g., images) for KB-VR tasks, due to the steep complexity of knowledge representation and processing in non-textual modalities.

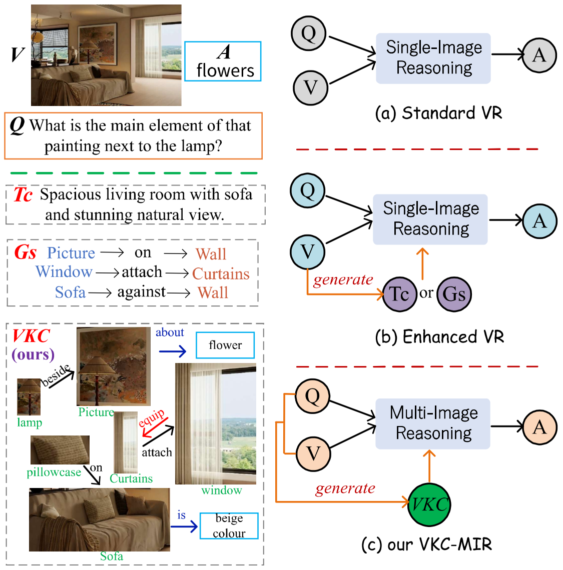


Figure 1: Motivation illustration. T_c, G_s, and VKC denote image caption, scene graph, and visual knowledge card (generated by our method). In our VKC-MIR framework, we employ a multi-image MLLM to address the proposed multi-image input <V,VKC,Q> (different from the standard input <V,Q> and existing enhanced inputs such as <V,Q,T_c/G_s>).

Advanced image-sequence understanding capabilities are required in practical applications, such as multi-image instruction (Jiang et al., 2024; Wu et al., 2024a). However, popular MLLMs insert visual features into the sequence of embeddings, which can easily exhaust the language model's context window, resulting in significant memory and computational overhead (Li et al., 2024b). Recent studies (Liu et al., 2024d; Li et al., 2024a)

[*]Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21883–21896
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

show that multi-image MLLMs (e.g., mPLUG-Owl3 (Ye et al., 2024)), which are pre-trained or fine-tuned on interleaved text-image data or multi-image data, perform better than single-image models (e.g., Qwen-VL-Chat (Bai et al., 2023)) when addressing multi-image inputs. However, if the images contain distracting content, the MLLMs will instead get confused, leading to a performance drop (Liu et al., 2024d). Therefore, providing relevant images that can convey task clues is beneficial for boosting the multi-image abilities of MLLMs.

Inspired by these advances on multi-image MLLMs, we aim to explore the new perspective of knowledge acquisition and representation for KB-VR and pose the question: *Could knowledge be described in the form of images so that we can employ a multi-image MLLM to solve an original single-image task enhanced by generating "knowledge image" as additional visual input?* To this end, as shown in Fig.1 (c), we introduce the novel concept named **V**isual **K**nowledge **I**mage (**VKC**), which aims to convey knowledge clues recognized for KB-VR tasks, and a corresponding framework named **VKC**-enhanced **M**ulti-**I**mage **R**easoning (**VKC-MIR**), which first generates VKC from the original task and then solves the enhanced task with the generated VKC using multi-image MLLMs. In particular, to decrease irrelevant knowledge, we first construct VKC as a coarse scene image composed of detected entity regions and then integrate it with external knowledge facts relevant to visual concepts in the image. Consequently, VKC-MIR consists of four stages: (1) visual scene perception; (2) external knowledge generation; (3) knowledge image editing; and (4) multi-image reasoning.

Our work opens new research directions and highlights the need for knowledge carriers to handle multi-modal knowledge challenges. **Our contributions** can be summarized as follows: (1) We introduce **the novel concept** of VKC, which presents knowledge in the form of images, offering a new perspective for multi-image MLLMs to handle KB-VR tasks. VKC also enjoys several advantages: (i) is a vivid image composed of key regions and knowledge details of a raw image; (ii) is task-independent, which can be applied to any visual task where raw images are provided; (iii) is model-agnostic, which can be applicable to a variety of multi-image models. (2) We propose **the novel framework**, VKC-MIR, which integrates the proposed VKC with MLLMs. The comprehensive four-stage pipeline, encompassing visual scene perception, external knowledge generation, knowledge image editing, and multi-image reasoning, provides a holistic solution to KB-VR. (3) Extensive experiments on popular benchmarks show that VKC-MIR achieves **state-of-the-art (SOTA)** results compared to the previous top models, showing its effectiveness and performance advantages.

2 Related Work

2.1 Knowledge-based Visual Reasoning

KB-VR requires the model to incorporate knowledge beyond the content of the given image and the question for answer prediction. Recently, integrating LLMs has significantly advanced state-of-the-art (SoTA) methods. PICa (Yang et al., 2022) incorporated GPT-3 for few-shot learning. Prophet (Shao et al., 2023) further refined the use of GPT-3 by prompting with answer heuristics. VCTP (Chen et al., 2024b) introduced visual Chain-of-Thought (CoT) prompting, which enhances KB-VR by guiding the model through a step-by-step reasoning process. KB-VR models often suffer from unwanted visual information that is not related to the question during retrieval. To this end, LLM-RA (Jian et al., 2024) identified key visual entities for multimodal joint retrieval. VIG (Liu et al., 2024e) introduced multi-grained visual information to retrieve knowledge. RZF-VQA (Wu et al., 2024c) reduced retrieval errors by associating external knowledge with a common feature space.

2.2 Multi-image MLLMs

Most previous MLLMs primarily handle single-image input scenarios, leaving the performance of MLLMs when addressing practical multiple images under-explored. To this end, Liu et al. (Liu et al., 2024d) proposed a new MIBench benchmark, to evaluate the multi-image abilities of MLLMs. The results revealed that, when faced with multi-image inputs, current models exhibited significant shortcomings, e.g., limited multi-image reasoning and fine-grained perception. Thus, MLLMs are suggested to be pre-trained or fine-tuned on interleaved image-text data. Idefics2 (8B) (Laurençon et al., 2024) performed a multi-stage pre-training using interleaved image-text documents, image-text pairs, and PDF documents. Unlike Idefics2 that collected noisy data from the Web, Mantis (8B) (Jiang et al., 2024) conducted instruction tuning on academic-level data resources. Integrating pre-training and

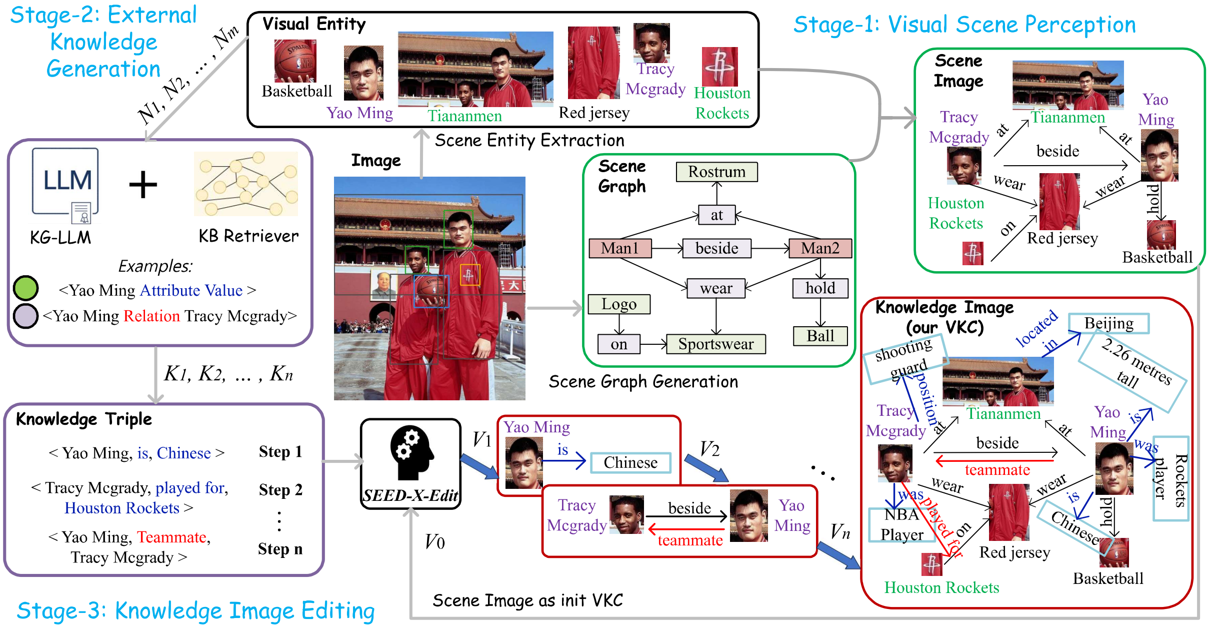


Figure 2: **VKC-MIR overview** (see Alg.1 for details). Stage-1: given an image, we first construct a scene graph using HiKER-SGG (Zhang et al., 2024), and then segment the global image into key entity regions using GLEE (Wu et al., 2024b), and subsequently visualize the scene graph by introducing entity regions using graphviz, thus generating our initial VKC. Stage-2: we employ OPT-66B (Zhang et al., 2022) as domain experts, to generate external knowledge about detected visual concepts. Stage-3: we perform message passing between LLM-generated knowledge and the updated VKC, using a text-to-image toolkit SEED-X (Ge et al., 2024b). Notably, for the space limit, we only exhibit the new message in each V_i here. Stage-4 (see Fig.1 (c)): we employ a multi-image MLLM, mPLUG-Owl3 (Ye et al., 2024), to solve our VKC-enhanced task.

instruction tuning on multi-image data, mPLUG-Owl3 (8B) (Ye et al., 2024) improved the ability to understand long image sequences.

In addition, we also introduce multi-modal knowledge graphs (MMKG) (Peng et al., 2023; Lee et al., 2024), which incorporate multi-modal knowledge (text, images, video, audio) in the large-scale graph structure, and discuss the distinctions between our concept and MMKG in Appendix F.

3 Methodology

Our method, VKC-MIR, is visualized in Fig.2. Given a raw image, VKC-MIR first generates a new scene image composed of detected entity regions and their spatial connections, and then integrates additional domain knowledge generated by LLM into scene image. *Two key benefits of our proposed staged approach are that*: (i) the generation task at each stage of our approach is simpler than learning to create a detailed knowledge image directly. Through each stage, the search space is restricted to the associated contents of the previously generated components, making VKC-MIR a powerful method for knowledge image generation; (ii) the data flow at each stage of our approach is essentially a collection of knowledge triples, which can be effectively integrated by advanced technologies such as entity linking or image editing. Making a fundamental determination whether or not a piece of knowledge should be incorporated into our knowledge image allows for more accurate answer predictions in the last stage of our approach. Next, we detail each stage in our VKC-MIR (Alg.1).

Algorithm 1 VKC-MIR Algorithm

Input: Input Image and Question $\{V, Q\}$.
Output: Visual Knowledge Card and Answer $\{VKC, A\}$.
Require: G_S denotes a scene graph for V; $E_M = \{E_1, ..., E_{|E_M|}\}$ is the set of visual entities detected in V and E_m is the m-th entity ($1 \leq m \leq |E_M|$); V_S denotes a scene image generated by visualizing G_S with E_M; $K_N = \{K_1, ..., K_{|K_N|}\}$ is the set of external knowledge generated by LLMs and K_n is the n-th knowledge ($1 \leq n \leq |K_N|$); V_n is the n-th knowledge image generated by introducing K_n; P_{kg}, P_{t2i}, and P_{mir} are the prompt for knowledge generation, image editing, and multi-image reasoning.

```
# Stage-1. Visual Scene Understanding
$G_S$ ← SceneGraphGenerator ($V$)
$E_M$ ← EntityExtractor ($V$)
$V_S$ ← SceneImageGenerator ($G_S$, $E_M$)
# Stage-2. External Knowledge Generation
Let $EN_M$ be entity names of $E_M$
$K_N$ ← LLMGenerate ($EN_M$, $Q$, $P_{kg}$) # See Alg.2
# Stage-3. Knowledge Image Editing
$V_0$ ← $V_S$
for $1 \leq n \leq |K_N|$ do
    $V_n$ ← ImageEditor ($V_{n-1}$, $K_n$, $P_{t2i}$)
$VKC$ ← $V_{|K_N|}$
# Stage-4. Multi-image Reasoning
$A$ ← MultiImageMLLM ($V$,$VKC$,$Q$, $P_{mir}$)
return $\{VKC, A\}$
```

Stage-1: Visual Scene Perception. Scene graph generation (SGG) offers a structured depiction of an image via the identification of objects and their relations. To effectively organize visual concepts in VKC, we first construct a scene graph G_S (Step 2) using a robust SGG engine, HiKER-SGG (Zhang et al., 2024), which exhibits the powerful abilities of generating scene graphs on both uncorrupted and corrupted images (e.g., containing real-world weather corruptions such as fog, snow, and smoke). The constructed G_S is a collection of textual triples $< E_h, SR, E_t >$ where E_h, SR, and E_t indicate head entity, spatial relation, and tail entity, respectively, e.g., <Man1, beside, Man2>. Then, we utilize GLEE (Wu et al., 2024b), to detect key visual entities and segment the global image into small entity region images E_M (Step 3). Next, we obtain a scene image V_S (Step 4), using a graph visualization software, graphviz[1]. Concretely, we adopt its layout programs that take textual descriptions of graphs and produce diagrams in the pre-defined formats, e.g., replacing each entity node E_h/E_t in G_S with its region image E'_h/E'_t and converting each SR node in G_S into a directed dependency edge that is labeled with SR. Consequently, V_S becomes a collection of multi-modal triples $< E'_h, SR, E'_t >$. Such directed edges also facilitate message passing, enabling the updating of knowledge representations in VKC.

Stage-2: External Knowledge Generation. Following (Chen et al., 2024a; Alamdari et al., 2024; Xu et al., 2024b), we employ LLM as a domain expert, to generate triplets of specialized and precise textual knowledge. Furthermore, to migrate knowledge hallucination (Liu et al., 2024b; Ji et al., 2023), our knowledge generation comprises two key components: *knowledge generator* and *knowledge verifier*. In summary, the generator first prompts the LLM to generate domain-specific knowledge for given entities. Subsequently, the verifier recognizes and filters the erroneous triples generated by the generator, to promote the precision of generated KGs. Concretely, the verifier uses existing criteria mined from open KGs within RuleHub (Ahmadi et al., 2020) to identify format errors and some conflict errors.

The above verifier can check basic contradictory knowledge, e.g., ensuring that a person's age is not a negative number. However, when facing complex knowledge conflicts (e.g., an NBA player belongs to different teams at different time periods), the verifier fails to work because conflicting tuples are correct historical facts. Therefore, we add a *history knowledge verifier* after this basic verifier, to explore better solutions to contradictory knowledge. We describe this in Appendix D due to space limit.

Algorithm 2 *LLMGenerate* Algorithm

Input: Entity Names, Question, and Prompt $\{EN_M, Q, P_{kg}\}$.
Output: Generated Knowledge $\{K_N\}$.
Require: D_m is domain knowledge generated by LLM given EN_m ($1 \leq m \leq |EN_M|$); T_m is a set of triple examples found by the retriever R_{KB} from knowledge base KB regarding EN_m; AK_m is a set of generated attribute knowledge regarding N_m; $NN_{<m,l>}$ is an entity-entity pair with EN_m ($1 \leq m \leq |EN_M|$) as head and EN_l ($1 \leq l \leq |EN_M|$) as tail; $D_{<m,l>}$ is domain knowledge generated by LLM given $NN_{<m,l>}$; $T_{<m,l>}$ is a set of triple examples found by R_{KB} from KB regarding $NN_{<m,l>}$; $OK_{<m,l>}$ is a set of generated object knowledge regarding $NN_{<m,l>}$; KC is the set of knowledge candidates and KC_v is the resulting set after using the verifier; P_{kg} is the prompt for knowledge generation.

```
$AK, OK \leftarrow \emptyset$
#Generate attribute knowledge
for $1 \leq m \leq |EN_M|$ do
    $D_m \leftarrow$ LLMGen $(EN_m)$
    $T_m \leftarrow$ KBR $(KB, EN_m)$
    $AK_m \leftarrow$ LLMGen $(D_m, T_m, P_{kg})$
    $AK \leftarrow AK \cup \{AK_m\}$
#Generate object knowledge
for $1 \leq m \leq |EN_M|$ do
    for $1 \leq l (\neq m) \leq |EN_M|$ do
        $NN_{<m,l>} \leftarrow < EN_m, EN_l >$
        $D_{<m,l>} \leftarrow$ LLMGen $(NN_{<m,l>})$
        $T_{<m,l>} \leftarrow$ KBR $(KB, NN_{<m,l>})$
        $OK_{<m,l>} \leftarrow$ LLMGen $(D_{<m,l>}, T_{<m,l>}, P_{kg})$
        $OK \leftarrow OK \cup \{OK_{<m,l>}\}$
#Verify by rules and re-rank by question $Q$
$KC \leftarrow AK \cup OK$
$KC_v \leftarrow$ RuleVerify $(KC)$
$K_N \leftarrow$ TupleRank$(KC_v, Q)$
return $K_N$
```

We elaborate the knowledge generation workflow in Alg.2 where entity names and the question are used as arguments. Each generated knowledge tuple depicts either attribute knowledge of a single entity or object knowledge between two entities detected from the raw image. We employ an open-source LLM, OPT-66B (Zhang et al., 2022), to generate the domain knowledge text D_m via a simple prompt (e.g., "Generate a paragraph of domain knowledge text (512 tokens) about the given entity"). Next, following (Chen et al., 2024a), we retrieve several reference triples T_m from DBpedia[2], an open-source encyclopedic KG, for few-shot prompting. Concretely, the LLM outputs two categories of knowledge: (i) for a specified entity

[1] https://graphviz.org/

[2] https://www.dbpedia.org/

EN_m, the LLM outputs attribute knowledge AK_m (line 6) in the form of $< EN_m, AR, E_t >$, where AR is a new attribute of EN_m learned from D_m and E_t is the corresponding attribute value; and (ii) for a specified entity-entity pair $< EN_m, EN_l >$, the LLM outputs object knowledge $OK_{<m,l>}$ (line 14) in the form of $< EN_m, OR, EN_l >$, where OR is a new object relation learned from $D_{<m,l>}$. In particular, we add D_m and T_m to P_{kg}, as exemplified in Fig.6 (Appendix C), where knowledge triples regarding "Yao Ming" are outputted. Next, we collect all generated tuple candidates (line 17).

After using the verifier to exclude erroneous tuples (line 18), we calculate the semantic similarity score for each "tuple-question" pair using Sentence-BERT (Reimers and Gurevych, 2019) (i.e., generating their sentence vectors and then calculating the cosine similarity between the pairwise vectors) and select the top tuples according to the resulting scores (line 19). Thus, *the selected knowledge tuples (or items) are relevant to the question.*

Stage-3: Knowledge Image Editing. VKC-MIR uses an image editing tool to generate the final image that looks like a vivid "knowledge card". Specifically, we employ SEED-X (Ge et al., 2024b), a text-to-image (T2I) toolkit as a follow-up work to SEED-LLaMA (Ge et al., 2024a). We design the template P_{t2i} below, which includes task instructions and editing requirements. To indicate data flow (e.g., through various instructions or manipulations), we mark each data element in blue.

> ********** Adding Attribute Knowledge **********
> **Instruction**: Strictly ensure that other elements of the image remain unchanged, adding the current knowledge $< EN_m$, attribute, value $>$ to the input image V_{n-1}.
> **Image Editing Requirements**: Add a box containing value next to EN_m and then point a blue arrow, attached the <attribute> name, from EN_m to that box.
> **Output**: V_n

> ********** Adding Object Knowledge **********
> **Instruction**: Strictly ensure that other elements of the image remain unchanged, adding the current knowledge $< EN_m$, relation, $EN_i >$ to the input image V_{n-1}.
> **Image Editing Requirements**: Add a red arrow from EN_m to EN_i and attach the <relation> to the arrow.
> **Output**: V_n

As shown in Steps 9-12 of Alg.1, we design an iterative process to gradually add the selected knowledge items to VKC rather than injecting all the items at once, which could lead to an erroneous image (see Sections 6.4 and 6.5). In addition, to prevent entity regions in VKC from being overcrowded, we use a string name rather than a random region image for each new entity. Given that P_{t2i} is followed, the final image will have high-level semantics that comply with these multimodal instructions and retain low-level details (e.g., the original entity regions retain uncorrupted).

Stage-4: Multi-image Reasoning. In the final stage, we use mPLUG-Owl3 (Ye et al., 2024) to solve the VKC-enhanced task. In particular, mPLUG-Owl3 uses Siglip-400m (Zhai et al., 2023) as visual encoder and Qwen2 (Yang et al., 2024) as language model. Given a multi-image input $<V, VKC, Q>$ in our enhanced task, mPLUG-Owl3 first extracts visual features for the image sequence $[V, VKC]$ and then uses a linear projection to align the dimensions of visual features to be the same as those of the language model. The projected visual features are denoted by $H_{img} = [I_V, I_{VKC}]$. Based on multi-modal input, the corresponding text sequence is $S_{text} = [T_{img}, T_{img}, Q]$, where T_{img} is a plain text $<|image|>$ to indicate the original location of the image. We feed S_{text} into a text encoder to obtain the text features H_{text}. In the language model, mPLUG-Owl3 then integrates H_{img} with H_{text} through cross-attention operations. In particular, we design a prompt template P_{mir} for mPLUG-Owl3 such as:"Look at Image 1 (VKC) which may provide knowledge clues and answer the following question (Q) based on Image 2 (V)".

4 Experiment Setup

4.1 Datasets and Compared Methods

We evaluate VKC-MIR on **three popular KB-VR benchmarks**: A-OKVQA (Schwenk et al., 2022), OK-VQA (Marino et al., 2019), and InfoSeek (Chen et al., 2023). Following (Chen et al., 2024b), we report the results (%) using the common VQA **accuracy** metric (Antol et al., 2015). We describe the details of datasets in Appendix A.

We compare VKC-MIR with four categories of existing methods: (1) Pure multi-modal methods (not using LLMs) that adopt uni-modal encoders to address the input of the corresponding modality, including **LXMERT** (Tan and Bansal, 2019), **KRISP** (Marino et al., 2021), **MAVEx** (Wu et al., 2022), **GPV-2** (Kamath et al., 2022), **UnifER** (Guo et al., 2022), and **RZF-VQA** (Wu et al., 2024c). (2) LLM-based methods that first generate descriptive captions about images and then employ language-only LLMs to address new textual input such that each image must be replaced with its caption, including **PICa** (Yang et al., 2022), **Prophet** (Shao et al., 2023), and **VCTP** (Chen et al., 2024b). (3) Open-source single-image MLLMs that improve the capabilities of LLMs by integrating data from multiple modalities, including **Qwen-VL-Chat** (Bai et al., 2023), **LLaVA-1.5** (Liu et al., 2024c), and **mPLUG-Owl** (Ye et al., 2023). (4) Open-source multi-image MLLMs that are pre-trained or fine-tuned on interleaved image-text data and multi-image data, including **Mantis** (Jiang et al., 2024), **Idefics2-I** (Laurençon et al., 2024), and **mPLUG-Owl3** (Ye et al., 2024). We describe the details of these methods in Appendix B. We did not widely compare with closed-source MLLMs such as GPT-4o, due to expensive APIs and limited visits.

4.2 Implementation Details

To control the quantity of knowledge items, we set implementation options below. In Stage-1, we set the maximum number of detected visual entities to 8. In Stage-2, we set the number of reference triples retrieved from DBPedia to 3. Also, we set the largest number of triples generated for a specific entity to 5 and the largest number of triples generated for an entity-entity pair (composed of two distinct entities) to 3. This may generate a total number of up to 8*5+8*7*3=208 candidates. In Stage-3, after parameter tests (Section 6.4), we set the total number of tuples (added to VKC) to 16 while the number added in each iteration to 1.

We performed all experiments on 4 NVIDIA 3090 24GB GPUs. Generating our VKC involves the interactions of multiple LLMs/MLLMs. Detailed computational costs are: (1) Memory: we keep the size of VKC consistent with that of the raw image, preventing a large increase in memory workload. For example, the average size of VKC on the OK-VQA dataset is only 1.72M. (2) Time: the inference time of the original/enhanced task is approximately 2.4/2.6 hours on average, showing that extra time of using VKC is negligible. (3) Parameter: the tools used in each stage are open-source and the whole pipeline does not require retraining on new datasets. Thus, *our method can be easily deployed for real-world applications.*

5 Results and Analysis

5.1 Quantitative Results

VKC-MIR exhibits superior performance than all other methods. Through the results in Table 1, we observe that: (a) VKC-MIR achieves the highest score of 58.9/64.8/25.1 on the A-OKVQA/OK-VQA/InfoSeek dataset, surpassing prior top methods such as Prophet (calling GPT-3) by 3.2/3.7/1.9 and VCTP (using Llama-2) by 4.5/9.9/3.7, respectively. (b) Pure multi-modal methods show much worse performance than LLM-based methods or MLLMs. (c) Among single-image MLLMs, Qwen-VL-Chat performs best and shows a performance comparable to that of Mantis. (d) The multi-image group performs better than the single-image group from an overall perspective. *Due to the excellent performance and availability of mPLUG-Owl3, we select it as our default MIR model in the last stage.*

Methods	A-OKVQA	OK-VQA	InfoSeek
Pure Multi-modal Methods (not using LLMs)			
LXMERT (2019)	25.9	35.3	6.7
KRISP (2021)	27.1	38.4	7.6
MAVEx (2022)	36.4	41.3	8.2
GPV-2 (2022)	40.7	39.7	8.6
UnifER (2022)	41.5	42.1	10.7
RZF-VQA (2024)	43.8	46.7	12.4
Language-only LLM-based Methods			
PICa-GPT3 API (2022)	47.1	48.0	16.9
Prophet-GPT3 API (2023)	<u>55.7</u>	<u>61.1</u>	23.2
VCTP (Llama2-70B) (2024)	54.4	54.9	21.4
Open-source Single-image MLLMs			
Qwen-VL-Chat (2023)	49.4	56.6	19.7
mPLUG-Owl (2023)	47.8	53.6	20.0
LLaVA-1.5 (2024)	44.6	47.3	17.2
Open-source Multi-image MLLMs			
Mantis (2024)	50.2	55.4	21.2
Idefics2-I (2024)	48.3	50.4	21.8
mPLUG-Owl3 (2024)	55.3	60.1	<u>23.4</u>
Our VKC-MIR *(using mPLUG-Owl3 (8B))*			
VKC-MIR	**58.9**	**64.8**	**25.1**

Table 1: VQA scores of different methods. We reproduced all other methods based on their released codes. The best/second-best result is highlighted in **bold**/<u>underlined</u> respectively.

Model	Base	+SG	+CAP	+VKC
open-source single-image MLLMs				
Qwen-VL-Chat	49.4	<u>50.8</u>	50.4	**52.0**
mPLUG-Owl	47.8	48.6	<u>48.9</u>	**50.2**
LLaVA-1.5	44.6	<u>46.5</u>	45.0	**46.8**
open-source multi-image MLLMs				
Mantis	50.2	<u>51.9</u>	51.3	**52.7**
Idefics2-I	48.3	<u>49.3</u>	48.9	**50.4**
mPLUG-Owl3	55.3	56.4	<u>56.5</u>	**58.9**

Table 2: Performance comparisons of distinct MLLMs, provided scene graphs (SG), image captions (CAP), and our VKC. The best/second-best result is highlighted in **bold**/<u>underlined</u>.

VKC can be applicable in various models and bring performance advantages against image captions and scene graphs. Concretely, we implement VKC-MIR via using three single-image MLLMs (Qwen-VL-Chat, mPLUG-Owl, LLaVA-1.5) and three multi-image MLLMs (Mantis, Idefics2-I, mPLUG-Owl3), respectively. In Table 2, we report the implementation results (VKC) using different models and also compare with their benchmark results (Base) and enhanced results with the given scene graphs (SG) or image captions (CAP) on the A-OKVQA dataset. As mentioned above, we use HiKER-SGG to generate SG and BLIP-2 to generate CAP for different models. Following (Chen et al., 2024b), we append SG/CAP at the end of the question text. From Table 2, we observe that: (1) For each MLLM, the "VKC" result is consistently better than the corresponding "Base" result, validating *the adaptability and model-agnosticism of our framework*; (2) Due to better capturing internal and external information, our VKC outperforms both CAP and SG; (3) As indicated by the "VKC" column, our VKC-MIR with mPLUG-Owl3 performs the best, showing the benefits of *using mPLUG-Owl3 as our default MLLM*.

5.2 Ablation Studies

To verify the contribution of key components, we examine six ablation variants. (i) Only MIR: we use the MIR model to solve the original task; (ii) G_S + MIR: we only generate scene graphs (a collection of textual triples) to enhance the input of MIR; (iii) K_e + MIR: we only generate external knowledge in the form of textual triples to enhance the input of MIR; (iv) G_S + K_e + MIR: we generate both scene graphs and external knowledge to enhance the input of MIR; (v) ($G_S \rightarrow V_S$) + MIR: we construct scene graphs and then visualize them to images (the initial VKC) to enhance the input of MIR; (vi) ($G_S \rightarrow V_S$) + K_e (T2I) + MIR: we implement all stages of VKC-MIR, thus generating our final model. Through the results in Table 3, we observe that: (1) Using G_S solely achieves better results than using K_e solely. However, using them jointly (G_S + K_e) can further improve the model's performance. (2) Converting G_S into images ($G_S \rightarrow V_S$) leads to performance gains compared to using G_S itself. This benefits from the multi-image abilities of MLLMs. (3) Our final (full) model achieves the best results compared to all tested variants, showing that each component contributes uniquely to the overall performance.

Methods	A-OKVQA	OK-VQA	InfoSeek
VKC-MIR (full)	**58.9**	**64.8**	**25.1**
Ablation variants (generated by gradually adding components)			
- only MIR model	55.3	60.1	23.4
- G_S + MIR	56.4	62.3	23.9
- K_e + MIR	56.0	61.9	23.6
- G_S + K_e + MIR	57.3	63.5	24.1
- ($G_S \rightarrow V_S$) + MIR	57.7	63.8	24.3
- ($G_S \rightarrow V_S$) + K_e (T2I) + MIR	**58.9**	**64.8**	**25.1**

Table 3: Ablation studies. ($G_S \rightarrow V_S$) denotes that scene graph (G_S) is converted to scene image (V_S) by graphviz. K_e (T2I) denotes that K_e is added to V_S through Text-to-Image.

5.3 Qualitative Results

We provide a qualitative comparison of VKC with common categories of image information in Fig.3. For fair comparisons, we uniformly use mPLUG-Owl3 to address enhanced inputs in various cases. Observed from Fig.3, CAP generates a summarized description about the image, while lacking enough attention to key visual concepts (e.g., "lamp" in Case 1) that are semantically important to the question. In contrast, SG captures the essential spatial relation between visual concepts (e.g., "<lamp, beside, picture>"), but it still gets confused due to the lack of fine-grained information that matches the visual content. Compared with them, our VKC not only adaptively attends to visual concepts which are semantically important to the question but also captures a trace of knowledge clues (e.g., "Yao Ming" (Entity 1) -> "Tracy Mcgrady" (Entity 2) -> "Houston Rockets" (guided by Relation "played for") in Case 2) during answer prediction.

6 Discussions

6.1 Comparisons with Closed-source MLLMs

We compare with mainstream large-scale closed-source MLLMs (e.g., GPT-4o). We perform this experiment on the val/test sets of the A-OKVQA dataset. For more comparisons, we also show the results of our default MIR model, mPLUG-Owl3. Observed from Fig.4, GPT-4o performs better than mPLUG-Owl3. However, our method, which generates VKC as an assisted visual input to mPLUG-Owl3, can surpass GPT-4o in the datasets studied.

6.2 Using Distinct LLMs as a Domain Expert

We present the results of using different LLMs as a domain expert during knowledge generation. Specifically, we consider two new options: (i) Llama2-70B (Touvron et al., 2023), which is an open-source LLM with size similar to OPT-66B, and (ii) GPT3-175B (Brown et al., 2020), which is a closed-source LLM of larger scale. We call the APIs to GPT3. As Table 4 shows, our VKC-MIR using OPT-66B can achieve the final results on a par with those obtained by using GPT3. However, our applications with OPT-66B enjoy extra advantages, e.g., free of charge and easy to deploy.

LLMs as domain expert	A-OKVQA	OK-VQA	InfoSeek
OPT-66B	**58.9**	64.8	**25.1**
Llama2-70B	58.0	63.9	24.7
GPT3-175B	58.7	**65.1**	**25.1**

Table 4: Results of using different LLMs.

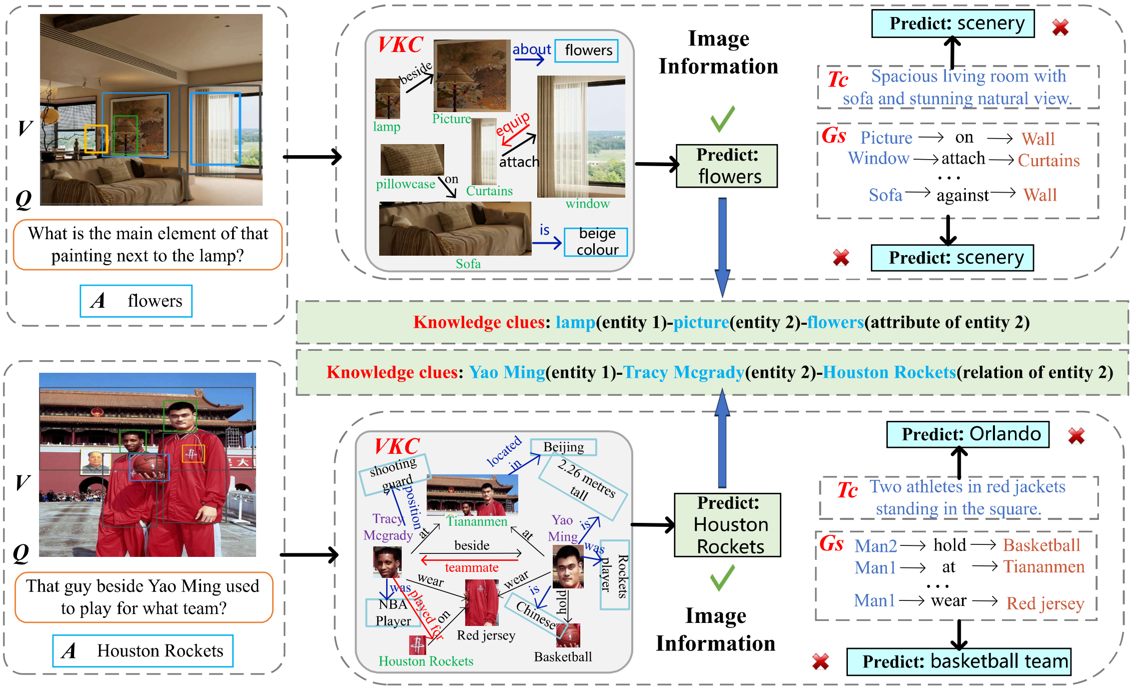


Figure 3: Qualitative results of using our VKC and other information categories such as image captions (T_C) and scene graphs (G_S). Our method also enjoys better interpretability by providing a trace of knowledge clues with related visual concepts.

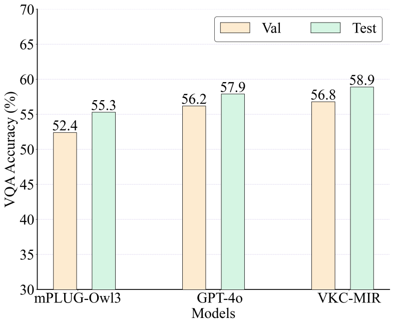


Figure 4: Comparisons with GPT-4o on the A-OKVQA.

6.3 The Effectiveness of Verifier and Ranker

In the second stage, we filter the generated knowledge candidates using a knowledge verifier and then rank the verified candidates using a knowledge ranker. To validate their effectiveness, we perform more fine-grained ablation studies, e.g., canceling one of these two components (w/o Verifier or Ranker). In particular, canceling the ranker indicates that the candidates are randomly selected. As shown in Table 5, our model performs best by jointly using the knowledge verifier and ranker, further validating the necessity of introducing these steps in our knowledge generation stage.

Method	A-OKVQA	OK-VQA	InfoSeek
VKC-MIR with Verifier + Ranker	**58.9**	**64.8**	**25.1**
VKC-MIR w/o Verifier	58.4	64.5	24.7
VKC-MIR w/o Ranker	58.1	64.3	24.5

Table 5: Results of using or canceling the verifier/ranker.

6.4 Super-parameter Test

In the third stage, we design an iterative process to gradually add the selected knowledge items to VKC. Specifically, we introduce two control variables: (i) the total number of knowledge triples $\#K_e$ added into VKC; and (ii) the number of knowledge triples $\#K_{e_i}$ added in each iteration. We compare the final results of using different quantities of these two variables, e.g., given $\#K_e \in \{4, 8, 16, 32\}$ and $\#K_{e_i} \in \{1, 2, 4, 8, 16\}$ ($\#K_{e_i} \leq \#K_e$). In Table 6, we find that: (1) Under the settings of $\#K_{e_i}$=1/2/4, the model's performance first rises and then drops if we introduce more and more knowledge (possibly sparse or irrelevant); (2) Under the settings of $\#K_{e_i}$=8/16, the model's performance continuously increases as $\#K_e$ grows, but their final results are inferior to those produced by setting a smaller value to $\#K_{e_i}$ (e.g., 1/2/4); (3) The best setting is considered as <$\#K_e$=16, $\#K_{e_i}$=1>, since we observe a highest accuracy 64.8%. These results show that we should add one new item each time, which is better than adding multiple items at the same time (possibly generating bad objects, as exemplified in Fig.5).

Parameters	$\#K_{e_i}$=1	$\#K_{e_i}$=2	$\#K_{e_i}$=4	$\#K_{e_i}$=8	$\#K_{e_i}$=16
$\#K_e = 4$	60.5	60.3	60.1	-	-
$\#K_e = 8$	62.5	62.2	61.5	*59.8*	-
$\#K_e = 16$	**64.8**	64.1	63.7	60.8	60.4
$\#K_e = 32$	63.9	63.4	62.0	61.2	60.6

Table 6: Parameter tests on the OK-VQA dataset.

6.5 Qualitative Analysis on VKC Quality

We provide a qualitative comparison of two similar VKCs generated by assigning the same value to $\#K_e$ but different values to $\#K_{e_i}$, given all other configurations are equal. As shown in Fig.5, referring to the left VKC, we find editing errors in the right VKC, such as missing relation annotations (e.g., "played for") and overlapping object boxes (e.g., "2.26 metres" and "Beijing"). This case provides explanations for the insights from Table 6.

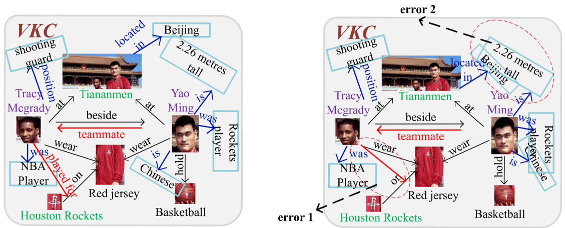


Figure 5: Comparisons of similar VKCs (Left: $\#K_{e_i} = 1$; Right: $\#K_{e_i} = 8$. Both: $\#K_e = 8$). Compared to the left VKC, we observe undesired editing effects including missing relation annotations (e.g., "played for") and overlapping object boxes (e.g., "2.26 metres tall" and "Beijing") in the right VKC.

6.6 The Necessity of Knowledge Images

To validate the necessity of knowledge images, we compare the results between generating VKC in the form of an image (which we call VKC-image) and converting all knowledge items in VKC back to textual triples (which we call VKC-triple). In Table 7, we observe that VKC-image outperforms VKC-triple in all datasets, suggesting the benefits of presenting vision knowledge in the form of images. Furthermore, we explore the impact of varying the number of VKCs (i.e., using multiple images or a single image) and find that: If knowledge triples are used to generate multiple VKCs, the model will perform worse instead. See Appendix E for details.

Knowledge Form	A-OKVQA	OK-VQA	InfoSeek
VKC-triple	57.3	63.5	24.1
VKC-image	**58.9**	**64.8**	**25.1**

Table 7: Comparison of VKC presented in triples/image.

6.7 Results on Multi-image Tasks

To explore multi-image applications of our proposal, we conduct additional experiments on four multi-image datasets released by MIBench (Liu et al., 2024d): (1) Subtle Difference (SD) task examines the model's ability to perceive subtle differences between similar images; (2) Visual Referring (VRef) task evaluates whether the model can utilize the referring information provided by input images to comprehend the relationships between different objects; (3) Vision-linked Textual Knowledge (VTK) task queries background knowledge that encompasses images and corresponding text which are possibly retrieved from a knowledge base (e.g., Wikipedia); (4) Text-linked Visual Knowledge (TVK) task needs the model to link the question to the relevant text, and extract visual information from the corresponding image. The results in Table 8 show that VKC-MIR can consistently outperform other MLLMs in various multi-image scenarios. In particular, in two knowledge-seeking tasks VTK and TVK, VKC-MIR achieves significant improvements compared to the others, due to its exploration of relevant external knowledge.

Methods	SD	VRef	VTK	TVK
		open-source single-image MLLMs		
Qwen-VL-Chat	22.5	16.3	22.9	18.1
mPLUG-Owl	4.0	21.7	14.9	20.6
LLaVA-1.5	14.9	24.1	16.7	26.3
		open-source multi-image MLLMs		
Mantis	54.1	37.6	26.4	41.7
Idefics2	49.7	32.6	25.6	39.0
mPLUG-Owl3	70.1	33.0	31.1	48.8
VKC-MIR (ours)	**72.5**	**42.6**	**40.7**	**55.2**

Table 8: Results of different methods in multi-image tasks.

7 Conclusions

Our work tackles the task of VQA which requires external knowledge beyond the information visible in the image, referred to as the KB-VR task. We propose the novel concept, Visual Knowledge Card (VKC), which is essentially a process that creates a new image by editing the scene graph (based on visual entities detected from the input image) with external knowledge obtained from LLM (considered as a domain expert). This new edited image, along with the original image and the question, is posed to MLLM which can take multiple images as input to produce the answer. Our proposed pipeline VKC-MIR achieves state-of-the-art results on the three KB-VR benchmarks (A-OKVQA, OK-VQA, and InfoSeek) and outperforms a variety of existing approaches, including multimodal methods, LLM-based approaches, and MLLMs, by a good margin. Future work includes: (i) We will extend our approach to other categories of vision-language tasks such as visual dialog (Liu et al., 2024a), which requires an agent to answer consecutive questions based on both visual content and dialogue history. (ii) We will explore more advances in the generation task at each stage to further improve the performance of our proposed pipeline.

Limitations

We summarize the limitations of this paper here. First, although we have taken measures to generate a wide range of knowledge tuples using the LLM, the level of generated tuples is still constrained. For example, relevant knowledge of "unseen entities", not detected from the original image, cannot be generated in our current algorithm yet. Therefore, we consider to introduce a new knowledge component, *the level counter*, which can record the level of knowledge tuples and thus decide the direction of generation. Specifically, the LLM takes new entities from each generated triple for this level as input and then proceeds to generate the next-level triples until a preset maximum level is achieved. Second, in essence, our concept VKC can be applied to a visual task that incorporates any number of input images. Real-world multimedia information, such as web pages and social media, generally contains multiple images and corresponding text in interleaved forms. Therefore, multi-image scenarios have greater practical value than single-image scenarios. In the future, we consider building *a novel benchmark* that incorporates both single-image tasks and multi-image tasks, offering a comprehensive evaluation platform to measure the performance of our VKC-assisted models in various visual reasoning scenarios.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62076092).

References

Naser Ahmadi, Thi-Thuy-Duyen Truong, Le-Hong-Mai Dao, Stefano Ortona, and Paolo Papotti. 2020. Rulehub: A public corpus of rules for knowledge graphs. *ACM J. Data Inf. Qual.*, 12(4):21:1–21:22.

Parand A. Alamdari, Yanshuai Cao, and Kevin H. Wilson. 2024. Jump starting bandits with llm-generated prior knowledge. In *Proc. EMNLP*, pages 19821–19833.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *Proc. ICCV*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proc. CVPR*, pages 18392–18402.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. NeurIPS*.

Markus J. Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn. Sci. Technol.*, 5(3):35083.

Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024a. SAC-KG: exploiting large language models as skilled automatic constructors for domain knowledge graphs. *CoRR*, abs/2410.02811.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proc. EMNLP*, pages 14948–14968.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proc. AAAI*, pages 1254–1262.

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

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024a. Making llama SEE and draw with SEED tokenizer. In *Proc. ICLR*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024b. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396.

Biao Gong, Shuai Tan, Yutong Feng, Xiaoying Xie, Yuyuan Li, Chaochao Chen, Kecheng Zheng, Yujun Shen, and Deli Zhao. 2024. Uknow: A unified knowledge protocol with multimodal knowledge graph datasets for reasoning and vision-language pretraining. In *Proc. NeurIPS*.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan S. Kankanhalli. 2022. A unified end-to-end retriever-reader framework for knowledge-based VQA. In *Proc. ACM MM*, pages 2061–2069.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for VQA. In *Proc. EMNLP*, pages 10939–10956.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483.

Zan-Xia Jin, Heran Wu, Chun Yang, Fang Zhou, Jingyan Qin, Lei Xiao, and Xu-Cheng Yin. 2023. Ruart: A novel text-centered solution for text-based visual question answering. *IEEE Trans. Multim.*, 25:1–12.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. In *Proc. ECCV*, pages 662–681.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. 2024. LLM4SGG: large language models for weakly supervised scene graph generation. In *Proc. CVPR*, pages 28306–28316.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *CoRR*, abs/2405.02246.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. In *Proc. ACL*, pages 10767–10782. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *CoRR*, abs/2404.16790.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Yuandi Li, Hui Ji, Fei Yu, Lechao Cheng, and Nan Che. 2025. Temporal multi-modal knowledge graph generation for link prediction. *Neural Networks*, 185:107108.

An-An Liu, Chenxi Huang, Ning Xu, Hongshuo Tian, Jing Liu, and Yongdong Zhang. 2024a. Counterfactual visual dialog: Robust commonsense knowledge learning from unbiased training. *IEEE Trans. Multim.*, 26:1639–1651.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proc. CVPR*, pages 26286–26296.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024d. Mibench: Evaluating multimodal large language models over multiple images. In *Proc. EMNLP*, pages 22417–22428.

Heng Liu, Boyue Wang, Yanfeng Sun, Xiaoyan Li, Yongli Hu, and Baocai Yin. 2024e. VIG: visual information-guided knowledge-based visual question answering. In *Proc. CSCWD*, pages 1086–1091.

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 *Proc. CVPR*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proc. CVPR*, pages 3195–3204.

Jinghui Peng, Xinyu Hu, Wenbo Huang, and Jian Yang. 2023. What is a multi-modal knowledge graph: A survey. *Big Data Res.*, 32:100380.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. EMNLP*, pages 3980–3990.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10674–10685.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proc. ECCV*, pages 146–162.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proc. CVPR*, pages 14974–14983.

Yaguang Song, Xiaoshan Yang, Yaowei Wang, and Changsheng Xu. 2024. Recovering generalization via pre-training-like knowledge distillation for out-of-distribution visual question answering. *IEEE Trans. Multim.*, 26:837–851.

Zhenqiang Su and Gang Gou. 2024. Knowledge enhancement and scene understanding for knowledge-based visual question answering. *Knowl. Inf. Syst.*, 66(3):2193–2208.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proc. EMNLP*, pages 5099–5110.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. *CoRR*, abs/2407.15754.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based VQA. In *Proc. AAAI*, pages 2712–2721.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. 2024b. General object foundation model for images and videos at scale. In *Proc. CVPR*, pages 3783–3795.

Sen Wu, Guoshuai Zhao, and Xueming Qian. 2024c. Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval. *IEEE Trans. Multim.*, 26:1790–1800.

Ning Xu, Zimu Lu, Hongshuo Tian, Rongbao Kang, Jinbo Cao, Yongdong Zhang, and An-An Liu. 2024a. Learning to supervise knowledge retrieval over a tree structure for visual question answering. *IEEE Trans. Multim.*, 26:6689–6700.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024b. Generate-on-graph: Treat LLM as both agent and KG for incomplete knowledge graph question answering. In *Proc. EMNLP*, pages 18410–18430.

An Yang, Baosong Yang, Binyuan Hui, and et al. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

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 *Proc. AAAI*, pages 3081–3089.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*, abs/2408.04840.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proc. ICCV*, pages 11941–11952.

Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia P. Sycara, and Yaqi Xie. 2024. Hiker-sgg: Hierarchical knowledge enhanced robust scene graph generation. In *Proc. CVPR*, pages 28233–28243.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

A Details of Experimental Datasets

OK-VQA (Marino et al., 2019) includes more than 14,000 questions that require reasoning or external knowledge to answer. The knowledge of OK-VQA covers a variety of knowledge categories such as science & technology, history, and sports. OK-VQA has 12,591 unique questions out of 14,055 total and 7,178 unique question words. Their images come from the COCO image dataset, so the dataset contains the same basic distribution of images. **Augmented OK-VQA** (A-OKVQA) (Schwenk et al., 2022) is composed of a diverse set of about 25K questions that require a broad base of common sense and world knowledge to answer. It provides multiple-choice as well as direct answer evaluation settings. There are 3 rationales associated to each question in the train set providing the explanation/knowledge for answering the question. The A-OKVQA dataset contains 24,903 (Question + Answer + Rationale) triplets in 17.1K (train) / 1.1K (val) / 6.7K (test) splits. **InfoSeek** (Chen et al., 2023) is tailored to incorporate information seeking questions that cannot be answered by common sense knowledge alone. The task tests the model's skill to discover knowledge in different specialized fields from images, providing in-depth knowledge and solving relevant practical problems. InfoSeek dataset contains 1.35 million samples.

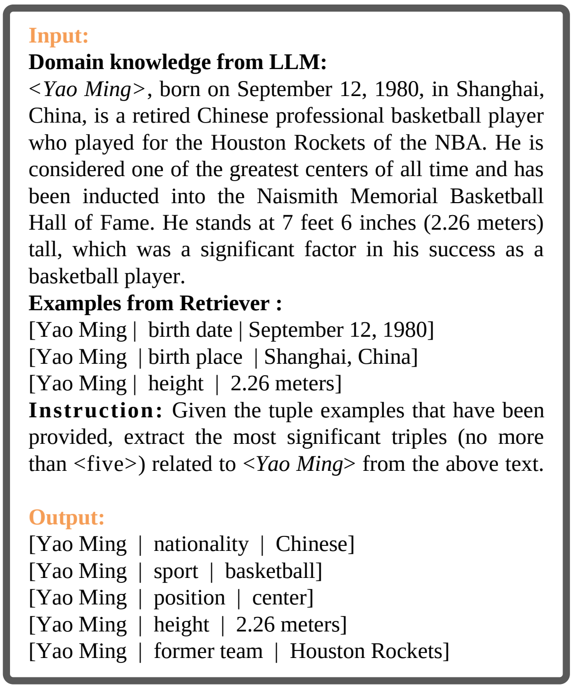
B Details of Compared Methods

In **Visual Reasoning (VR)**, numerous methods have been proposed to improve the performance of models in understanding and predicting questions related to visual content. LXMERT (Tan and Bansal, 2019) was a foundational model that employed a transformer-based architecture to encode visual and textual information. KRISP (Marino et al., 2021) improved LXMERT capabilities by integrating implicit and symbolic knowledge. MAVEx (Wu et al., 2022) introduced a multi-modal answer validation approach for KB-VR. GPV-2 (Kamath et al., 2022) leveraged web-supervised concept expansion to enhance general-purpose vision models. UnifER (Guo et al., 2022) presented a unified end-to-end retriever-reader framework for KB-VR. Recently, integrating LLMs has significantly advanced state-of-the-art methods. PICa (Yang et al., 2022) incorporated LLMs into VQA, using GPT-3 for few-shot learning. Prophet (Shao et al., 2023) further refined the use of GPT-3 by prompting with answer heuristics. VCTP (Chen et al., 2024b) introduced Visual CoT Prompting, which enhances KB-VR by guiding the model through a step-by-step reasoning process.

Multimodal Large Language Models (MLLMs) have been used in reasoning for vision-language tasks. On the one hand, recent studies have expanded the capabilities of LLMs to encompass multimodal contexts (Ye et al., 2024; Dai et al., 2023). Some research has delved into the enhancement of MLLMs with multi-image comprehension skills (Jiang et al., 2024; Laurençon et al., 2024). On the other hand, advanced diffusion models excel in generating high-quality images from textual prompts (Brooks et al., 2023; Rombach et al., 2022), while achieving the desired semantic content often necessitates multiple iterative attempts. Researchers are leveraging LLMs to enrich input prompts with contextual knowledge to enhance control over image generation.

C Details of Prompt Template P_{kg}

In Stage-2, we employ the open source OPT-66B to generate domain knowledge $D_m/D_{<m,i>}$ about an entity N_m/entity pair $NN_{<m,i>}$. Meanwhile, we follow (Chen et al., 2024a) to retrieve relevant triple examples $T_m/T_{<m,i>}$ from DBpedia for few-shot prompting to facilitate better knowledge generation. Thus, the prompt template P_{kg} is composed of $D_m/D_{<m,i>}$, $T_m/T_{<m,i>}$, and task description. Following this, we exhibit a detailed example in Fig.6, e.g., obtaining the attribute knowledge of a specific entity "Yao Ming".



Input:
Domain knowledge from LLM:
<*Yao Ming*>, born on September 12, 1980, in Shanghai, China, is a retired Chinese professional basketball player who played for the Houston Rockets of the NBA. He is considered one of the greatest centers of all time and has been inducted into the Naismith Memorial Basketball Hall of Fame. He stands at 7 feet 6 inches (2.26 meters) tall, which was a significant factor in his success as a basketball player.
Examples from Retriever :
[Yao Ming | birth date | September 12, 1980]
[Yao Ming | birth place | Shanghai, China]
[Yao Ming | height | 2.26 meters]
Instruction: Given the tuple examples that have been provided, extract the most significant triples (no more than <five>) related to <*Yao Ming*> from the above text.

Output:
[Yao Ming | nationality | Chinese]
[Yao Ming | sport | basketball]
[Yao Ming | position | center]
[Yao Ming | height | 2.26 meters]
[Yao Ming | former team | Houston Rockets]

Figure 6: Prompting examples for P_{kg}.

D Exploring Knowledge Conflicts

It is possible to introduce contradictory knowledge during our knowledge image generation. For instance, the NBA player Mcgrady played for different teams at different time periods (e.g., Orlando Magic (2000-2004) and Houston Rockets (2004-2010)). To this end, we explore a two-step validation approach to better address contradictory knowledge. In the first step, we still use the knowledge verifier introduced in the main text. The verifier is rule-based and parameter-free, enabling efficient error detection and correction. We use the verifier to check basic contradictory knowledge (BCK). Thus, we name the verifier, Rule-Based Validator for BCK (RV-BCK).

However, when facing conflicting tuples such as <Mcgrady, played for, Houston Rockets> and <Mcgrady, played for, Orlando Magic>, RV-BCK fails to work, because both knowledge items are correct historical facts. Therefore, in the second step, we employ an MLLM (e.g., mPLUG-Owl3) as a specific verifier of historical contradictory knowledge (HCK). We name the second verifier, MLLM-Based Validator for HCK (MV-HCK). Specifically, we feed the collection of knowledge tuples generated by the LLM, accompanied by the given image and the query question, as input to the MLLM. We require the MLLM to determine whether a specific knowledge tuple belongs to a past or current fact, conditioned on the given image and question. We try to use guiding prompts, such as "Focus on the given image and question...". Through initial experiments, we observe that MV-HCK can play an auxiliary role to RV-BCK. For instance, the knowledge tuple <Mcgrady, played for, Orlando Magic> can be kicked out through MV-HCK, as it is a past fact and not matched with the current question. We examine this on three datasets. The results in Table 9 show that the two-step validation method (RV-BCK + MV-HCK) can further enhance the performance of VKC-MIR.

Methods	A-OKVQA	OK-VQA	InfoSeek
VKC-MIR	58.4	64.5	24.7
VKC-MIR with RV-BCK	58.9	64.8	25.1
VKC-MIR with RV-BCK + MV-HCK	**59.3**	**65.2**	**25.3**

Table 9: Results of the two-step knowledge validation.

E The Impact of Varied Number of VKCs

We compare the effectiveness of multiple images and a single image to explore the impact of varying the number of knowledge images in this task. Specifically, given $\#K_e = 16$, we generate VKC by varying the number of images: (i) VKC(1), with all triples integrated into a single VKC; (ii) VKC(2), with the first eight triples integrated into the first VKC and the last eight triples integrated into the second VKC; (iii) VKC(4), with every four triples sequentially integrated into a new VKC. We perform this experiment on all three datasets. The results in Table 10 show that VKC(2) achieves a minor improvement compared to VKC(1), while VKC(4) shows a performance drop. This indicates that all knowledge terms should be added into a single image, ensuring the integrity of the information.

VKC Number Test	A-OKVQA	OK-VQA	InfoSeek
VKC(1)	58.4	**64.5**	24.7
VKC(2)	**58.7**	**64.5**	**24.9**
VKC(4)	58.1	64.2	24.3

Table 10: Comparison of the varied number of images.

F Multi-modal Knowledge Graph

MMKG (Peng et al., 2023; Buehler, 2024) can be expressed with text, images, video, audio, etc., extracting a variety of knowledge of modal entities such as elements, association of elements, and alignment. In essence, it is a relational data connection mode in the form of a semantic network (Lee et al., 2024). For instance, TMMKG (Li et al., 2025) generates a multi-modal temporal knowledge graph for link prediction. UKnow (Gong et al., 2024) introduces the unified Knowledge Protocol with MMKG datasets for reasoning and vision-language pre-training. Two key distinctions between MMKG and our VKC are: (i) MMKG is normally a super-size graph structure, incorporating multi-modal knowledge regarding a specific domain, while our VKC is a small-size realistic image generated to enhance the original task, incorporating internal visual knowledge that represents entity association and external world knowledge that details visual concepts; (ii) MMKG allows for a diverse range of multi-modal data (text, image, video, audio), while our VKC only allows for text and image (e.g., entity region images).