Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072903

ARTICLE

Metacognition Inspired Reflective Chain-of-Thought for Knowledge-Based VQA

Zhongfan Sun, Kan Guo, Yongli Hu* and Yong Zhang

School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China

*Corresponding Author: Yongli Hu. Email: huyongli@bjut.edu.cn

Received: 06 September 2025; Accepted: 08 December 2025; Published: 10 February 2026

ABSTRACT: Knowledge-based Visual Question Answering (VQA) requires the integration of visual information with external knowledge reasoning. Existing approaches typically retrieve information from external corpora and rely on pretrained language models for reasoning. However, their performance is often hindered by the limited capabilities of retrievers and the constrained size of knowledge bases. Moreover, relying on image captions to bridge the modal gap between visual and language modalities can lead to the omission of critical visual details. To address these limitations, we propose the **Re**flective **C**hain-**o**f-**T**hought (**ReCoT**) method, a simple yet effective framework inspired by metacognition theory. ReCoT effectively activates the reasoning capabilities of Multimodal Large Language Models (MLLMs), providing essential visual and knowledge cues required to solve complex visual questions. It simulates a metacognitive reasoning process that encompasses monitoring, reflection, and correction. Specifically, in the initial generation stage, an MLLM produces a preliminary answer that serves as the model's initial cognitive output. During the reflective reasoning stage, this answer is critically examined to generate a reflective rationale that integrates key visual evidence and relevant knowledge. In the final refinement stage, a smaller language model leverages this rationale to revise the initial prediction, resulting in a more accurate final answer. By harnessing the strengths of MLLMs in visual and knowledge grounding, ReCoT enables smaller language models to reason effectively without dependence on image captions or external knowledge bases. Experimental results demonstrate that ReCoT achieves substantial performance improvements, outperforming state-of-the-art methods by 2.26% on OK-VQA and 5.8% on A-OKVQA.

KEYWORDS: Knowledge-based VQA; metacognition; reflective chain-of-thought; answer refinement

1 Introduction

The real world is inherently multimodal, rich in both visual and linguistic signals. A core goal of Artificial Intelligence (AI) is to develop models capable of replicating human-like visual-linguistic understanding and causal reasoning [1,2]. Recent studies have emphasized that achieving such human-like intelligence requires integrating perception, reasoning, and compositional learning, enabling models to build structured causal representations that mirror human cognition [3,4]. Visual Question Answering (VQA) has emerged as a key benchmark for this goal. In particular, knowledge-based VQA extends the challenge by requiring models to integrate visual content with external open-domain knowledge, making it substantially more difficult than conventional VQA. By more closely reflecting human reasoning processes, it has become a foundational task in multimodal AI research.

Existing retrieval-based methods [5–7] generally acquire relevant information from external knowledge sources such as Wikipedia [8], ConceptNet [9], or Google Search [10], and incorporate it, along with the question and visual content, into pretrained models [9,11–13] for fine-tuning. The characteristics of

external knowledge reasoning enable Language Models (LMs) to outperform vision-language models, thereby becoming mainstream for subsequent research. However, due to limitations in the scale of knowledge bases or the capabilities of retrievers, the knowledge retrieved by these methods may be irrelevant or insufficient to solve complex visual problems. Moreover, many existing approaches address the modality gap between image and text by converting images into text. LMs commonly rely on image captions as a textual proxy for visual input, which may cause them to overlook the critical visual elements actually required to answer the question. For example, as illustrated in Fig. 1, given the question "*What material is burning?*", both the image caption and the retrieved knowledge focus on the more salient object, the *umbrella*, rather than the actual burning *candle*. Relying solely on these cues makes it difficult for the model to predict the correct answer, *wax*, thereby highlighting the limitations of caption-based grounding and external retrieval in complex visual reasoning scenarios.

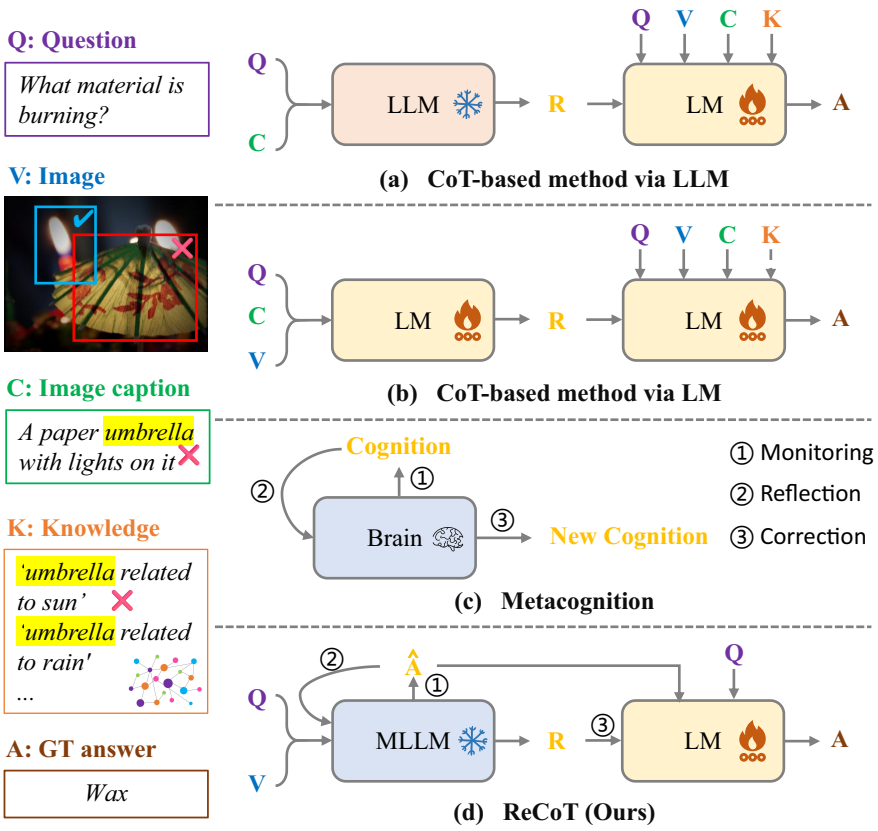

Figure 1: An overview of different Chain-of-Thought (CoT) based reasoning paradigms for knowledge-based VQA. (**a**) CoT-based methods via Large Language Models (LLMs) generate rationales based on image captions, followed by answer generation through a smaller Language Model (LM). (**b**) CoT-based methods via LM fine-tune smaller LMs with multimodal inputs to generate rationales. (**c**) Illustration of the metacognitive reasoning process, involving monitoring, reflection, and correction. (**d**) The proposed ReCoT framework simulates the metacognitive reasoning process by using a Multimodal Large Language Model (MLLM) to generate a preliminary answer and reflective rationale, which is then refined by a smaller LM to produce the final answer

To address the shortcomings of existing methods, our goal is to elicit large models to generate effective Chain-of-Thought (CoT) rationales that encapsulate the essential knowledge and visual cues needed to answer complex visual questions, without relying on external knowledge retrieval or image captions. Mimicking the procedure of human thinking and reasoning, the CoT-based approaches improve

question-answering accuracy and interpretability by generating CoT rationales to infer answers. Some methods [14–17] convert visual images into captions and feed them into frozen Large Language Models (LLMs) [18] to generate CoT rationales (as shown in Fig. 1a). Although the CoT rationales generated by these methods contain valuable knowledge from LLMs, they still require external knowledge retrieval to enhance performance in small model reasoning [14,15,17] and are prone to missing critical visual details. To better integrate visual information, recent multimodal CoT-based methods such as MM-CoT [19], T-SciQ [20] and PS-CoT-Adapter [21] have been applied to datasets like ScienceQA [22]. These approaches fine-tune smaller language models with visual features to generate detailed CoT rationales, achieving strong performance where all necessary information is available. However, ScienceQA provides complete contextual knowledge, making such methods effective only when reasoning does not require external information. In contrast, knowledge-based VQA tasks rely heavily on external factual knowledge beyond the image. Generating rich and accurate CoT rationales with small fine-tuned models is therefore highly challenging, as further evidenced in Section 4.4. Additional methods [23,24] have adopted more complex reasoning processes to improve performance. However, these approaches lack theoretical guidance, and the manually designed reasoning chains often fail to provide a simple and effective framework.

Although CoT-based methods emulate certain aspects of human thinking and reasoning, they remain far from replicating the cognitive capabilities of human intelligence. According to the metacognition theory [25,26], to solve the complex problems, humans generally have two cognitive levels: the basic cognition and the metacognition. The former helps us quickly acquire information and make preliminary judgments, while the latter metacognition cognizes the basic cognitive process. The metacognition includes monitoring, reflection, and correction of basic cognition. Especially, the reflective reasoning mechanism of metacognition can reflect the basic cognitive results, thereby enabling more flexible and effective solutions to complex tasks. Traditional cognitive reasoning in VQA models focuses primarily on direct perception and reasoning without evaluating the reliability of the outputs. These models rely mainly on feed-forward reasoning processes that lack higher-order self-monitoring and error correction. However, complex knowledge-based VQA tasks often require evidence validation and error correction—capabilities that standard reasoning processes do not possess. To address these challenges, it is essential to integrate human-like metacognitive mechanisms into artificial reasoning frameworks, bridging human cognitive reflection and machine inference. Based on this idea, we propose the **Re**flective **C**hain-**o**f-**T**hought (**ReCoT**) me thod to simulate the human metacognitive process by allowing the machine to monitor, reflect, and correct during reasoning to optimize the results. Specifically, as shown in Fig. 1c,d, drawing inspiration from the theoretical models [25,27], our proposed ReCoT method simulates the human metacognitive cycle through three corresponding stages: (1) generates the preliminary answer, similar to the monitoring process in metacognition; (2) generates the underlying reflective rationale based on the preliminary answer, akin to the reflection process in metacognition; and (3) refines the preliminary answer using the rationale, analogous to the correction process in metacognition. From the perspective of cognitive psychology, this process aligns with the Dual-Process Theory [28] and Cognitive Load Theory [29], where initial reasoning corresponds to the fast, intuitive System 1, and reflective reasoning corresponds to the slow, deliberate System 2. By enabling interaction between these two modes, ReCoT effectively balances cognitive efficiency and accuracy, reducing reasoning load while enhancing decision robustness.

Unlike traditional CoT or self-reflection approaches, which often rely on manually designed processes and complex model architectures, ReCoT introduces a significant innovation by drawing inspiration from metacognition theory. ReCoT goes beyond these methods by simulating a metacognitive process—specifically the stages of monitoring, reflection, and correction. An MLLM serves as the agent for monitoring and reflection, while a fully trained smaller language model is used for flexible correction. This compact

approach enables high-quality answer generation without the need for additional resources such as knowledge retrieval or image captions, surpassing current state-of-the-art performance. Furthermore, beyond benchmark performance, ReCoT's metacognitive design demonstrates potential for broader social impact. In educational technology, it can model self-reflective learning strategies; in medical image interpretation, it can support diagnostic reasoning with transparent reflection steps; and in complex decision-support systems, it can enhance interpretability and reliability. These applications illustrate the generalizability and societal relevance of metacognition-inspired AI reasoning.

Our contributions include the following three aspects:

- We propose a simple yet effective three-step CoT method, ReCoT, consistent with metacognition theory that incorporates a reflective reasoning approach, where answers and rationales mutually reinforce each other to enhance overall performance.
- The preliminary answers and reflective CoT rationales generated by our approach encapsulate the essential visual elements and domain-specific knowledge necessary for addressing visual questions. With these in place, accurate question answering can be accomplished using only a compact LM, eliminating the need for external knowledge retrieval or image captions.
- Experimental results demonstrate the effectiveness of our approach, surpassing the state-of-the-art methods by 2.26% and 5.8% on OK-VQA and A-OKVQA, respectively.

The remainder of this paper is organized as follows. Section 2 reviews the relevant research in knowledge-based VQA. Section 3 introduces the proposed ReCoT framework. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper by summarizing the key findings and outlining directions for future research.

2 Related Work

In this section, we review related work from three perspectives. First, we summarize recent advances in MLLMs, which refer to models trained on large-scale image-text data to enable cross-modal reasoning. These models facilitate visual-language reasoning by integrating visual encoders with language models. Then, we introduce CoT techniques, which focus on reasoning frameworks that break down complex problems into intermediate steps, thereby enhancing multi-step reasoning abilities. Finally, we discuss knowledge-based VQA methods, which extend traditional VQA by incorporating external knowledge or leveraging internal knowledge within models to answer more complex queries.

2.1 Multimodal Large Language Models

MLLMs [30–33] have been substantially advanced by large-scale image-text paired pre-training [34–36], significantly enhancing their capabilities in image-language understanding and reasoning. These models typically consist of a pretrained image encoder and a pretrained LLM. For instance, Alayrac et al. [30] employ gated cross-attention blocks to establish connections between the image encoder and the LLM. In contrast, Liu et al. [33,37] directly project visual features into the text embedding space through a multilayer perceptron. Li et al. [31] introduce the Q-Former module, which effectively narrows the gap between modalities. Earlier multimodal pretraining frameworks such as UNITER [38] laid the foundation for unified visual-text representation, which greatly influenced the subsequent development of MLLMs. Recently, to improve feedback optimization for MLLMs, Li et al. [39] leverage GPT-4V to generate high-quality feedback for feedback optimization in MLLMs. To address the issue of error correction simulation in MLLMs, reference [40] collects explainable feedback and provides new distractors for guiding the model.

However, unlike these works, we focus on the limitations in knowledge retrieval and the quality of CoT-based answer generation. We propose leveraging metacognitive theory to generate reflective CoT in large models to refine preliminary answers.

2.2 CoT Techniques

LLMs [18] have recently demonstrated exceptional performance, particularly in complex problem-solving, by employing CoT techniques to enhance multi-step reasoning capabilities [41]. Two major paradigms of CoT techniques have been developed: Zero-Shot-CoT [42], which prompts LLMs to reason step-by-step without demonstrations, and Few-Shot-CoT [41,43], which leverages annotated exemplars to guide reasoning. The evolution of CoT can be traced back to intermediate reasoning models [44], which first introduced explicit intermediate computation steps to improve transparency and reasoning control, paving the way for modern CoT approaches. Beyond single-pass reasoning, recent research has explored multi-round CoT reasoning [45,46], where models iteratively refine and verify their intermediate thoughts across reasoning turns, simulating deeper metacognitive reflection. These developments lay the groundwork for scalable reflective reasoning frameworks like ReCoT. Recent studies [19,20,47] have investigated multimodal CoT reasoning in multimodal scenarios by employing fine-tuned LMs to enhance their performance. Zhang et al. [19] initially introduces the concept of multimodal CoT, separating the processes of rationale generation and answer inference. Mondal et al. [47] enhance rationale generation and answer inference through the support of knowledge graphs, achieving greater effectiveness. Wang et al. [20] leverage LLMs to generate high-quality CoT rationales as teaching signals, addressing the disadvantages of costly human-annotated rationales. However, applying the aforementioned multi-modal CoT strategy to knowledge-based VQA tasks does not yield the expected results. Our proposed method generates high-quality rationales by introducing a simple yet effective framework, which significantly improves the final performance.

2.3 Knowledge-Based VQA

In the broader field of VQA, recent advances have explored mechanisms to overcome language priors and enhance compositional reasoning in multimodal understanding. For instance, Chowdhury and Soni [48] mitigate the language prior problem by enhancing visual feature representation through an ensemble of spatial and channel attention mechanisms, which improves visual–text fusion and reduces bias toward frequent answers. Building on this direction, Chowdhury and Soni [49] propose a unified framework that jointly tackles language prior and compositional reasoning challenges, enabling answer generation beyond a predefined answer space and improving overall robustness across VQA benchmarks. In various visual reasoning tasks, external knowledge is crucial for integrating additional semantic information and enhancing inference capabilities [50–52]. Early knowledge-based VQA methods retrieve knowledge from various external sources and integrate the retrieved knowledge into pretrained models for inferring answers [5,53,54]. Retrieval methods have evolved from traditional BM25 (Best Matching 25) to Dense Passage Retrieval (DPR) [55], and more recently to multimodal selection [56]. The progressively stronger retrieval capabilities have further enhanced the performance of knowledge-based VQA models. However, persistent technical bottlenecks remain—such as knowledge retrieval failure (when relevant information is absent or inaccessible) and semantic drift (when the reasoning chain diverges from the visual-linguistic context). These issues often lead to inconsistent or hallucinated answers, underscoring the need for robust self-monitoring mechanisms. ReCoT addresses these challenges by embedding reflective reasoning and correction, dynamically calibrating reasoning paths to maintain semantic coherence. Due to the nature of external knowledge reasoning, natural language pretrained models achieve better performance and have become mainstream [6,8,52]. Lin and Byrne [6] utilize DPR to search for knowledge from knowledge corpus

collected from [10] and infer answers in an end-to-end manner. Lin et al. [57] enhance the knowledge retrieval capabilities of [6], addressing the issue of low granularity in image and text representations when retrieving external knowledge. Hu et al. [52] encode various multimodal knowledge sources into large-scale memory, resulting in significant performance improvements. However, due to limitations in the retriever's capabilities or the scales of the external knowledge bases, the model may fail to retrieve the necessary knowledge to answer questions. Additionally, recent works on visual-language alignment [31] and context-aware reasoning [23] have further advanced multimodal understanding. These studies emphasize aligning fine-grained visual cues with linguistic semantics and reasoning adaptively based on contextual feedback. Building upon these advances, ReCoT complements such alignment-based reasoning with a metacognitive reflective process that continuously monitors and refines its visual-linguistic understanding.

Recent works [23,58–60] have attempted to leverage the implicit knowledge within large models to mitigate the limitations of external knowledge retrieval. These efforts suggest that large models could potentially serve as a replacement for external retrieval. However, as large models accumulate more knowledge, they tend to lose precision when answering specific questions, posing a new challenge in how to effectively harness their internal knowledge. To address this, recent studies [20,23,45] have explored strategies for triggering large model knowledge during CoT generation. In addition, Wang and Ge [61] have been proposed to mine visual cues from images and generate question-answer pairs to guide MLLMs in enhancing their reasoning capabilities. Although these approaches attempt to activate large model knowledge, they have yet to fully exploit the knowledge embedded within large models. In contrast, our approach leverages reflective thinking mechanisms inspired by metacognition to more effectively activate and utilize large model knowledge for precise and contextually grounded reasoning.

3 Proposed Method

As shown in Fig. 2, ReCoT simulates a metacognitive process—specifically the stages of monitoring, reflection, and correction—which correspond to the stages of Preliminary Generation, Reflective Reasoning, and Answer Refinement, respectively [25,27]. In the following, these three components are depicted in detail.

3.1 Initial Generation

Due to the complexity of questions and the requirement for rationales to incorporate extensive visual and commonsense knowledge, directly generating effective rationales is highly challenging. To address this, inspired by reflective thinking in metacognitive theory, we first generate a preliminary answer, which serves as the basis for subsequent reflective rationale generation. This differs from previous approaches that directly generate rationales.

Specifically, we use a frozen MLLM to generate initial responses to complex questions. We denote the VQA dataset as $D = (V_i, Q_i, A_i)_{i=1}^{N}$, where V_i, Q_i, and A_i represent the image, question, and a list of ten ground-truth (GT) answers, respectively. The following instruction is constructed to acquire the preliminary answer of the MLLM for each sample in each split:

> Ins_1: Answer the question using a single word or phrase.

The preliminary answer a_{MLLM} is generated by feeding the instruction Ins_1, image V, and question Q into the MLLM. The specific formula is as follows:

$$a_{MLLM} = f_{MLLM}(Ins_1, V, Q), \tag{1}$$

where f_{MLLM} refers to the frozen MLLM. To generate more accurate preliminary answers, we adopt LLaVA as the MLLM, as it is specifically trained for multimodal VQA tasks.

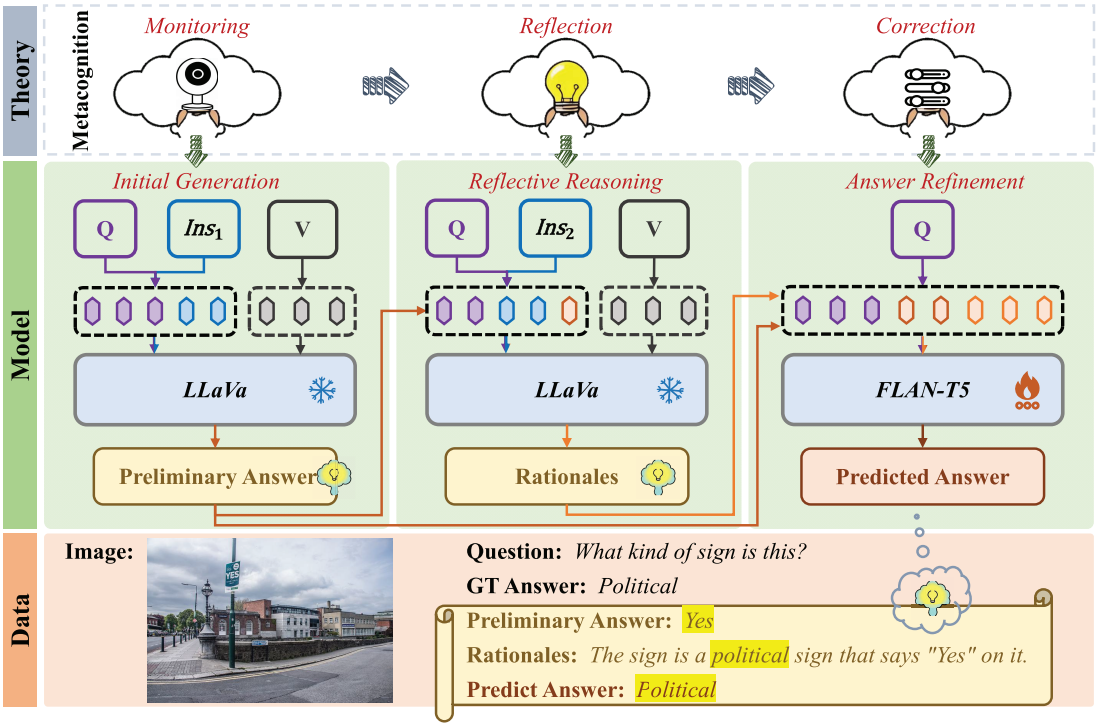

Figure 2: The proposed ReCoT method. The upper, middle, and lower parts display Theory, Model, and Data, respectively. The model is divided into three main steps: initial generation, reflective reasoning, and answer refinement, which correspond to the monitoring, reflection, and correction processes in metacognitive theory

Although simple in structure, the Initial Generation stage encapsulates the model's comprehension of both visual and textual inputs, producing a preliminary answer that represents its cognitive state. This output serves as the foundation for subsequent meta-level monitoring in the Reflective Reasoning stage, consistent with the object-level cognition defined in metacognitive theory [27].

3.2 Reflective Reasoning

The CoT rationale enhances the interpretability and performance of the model. To generate a rationale that incorporates both visual and knowledge information relevant to the question, we reflect on the preliminary answer to gain a deeper understanding of the initial reasoning.

However, unlike previous methods that use image captions to bridge the modal gap between images and LMs, which leads to the loss of visual information, we use LLaVA that can directly process images to generate rationales. Specifically, we construct the following instruction:

> Ins_2: Please explain the answer briefly with one short paragraph.

The CoT rationale $R_{a_{MLLM}}$ can be generated by combining the instruction Ins_2, image V, question Q and answer a_{MLLM} as inputs to the MLLM. The specific formula is as follows:

$$R_{a_{MLLM}} = f_{MLLM}(Ins_2, V, Q, a_{MLLM}). \tag{2}$$

This step is similar to the reflective regulation process in metacognition, which aims to improve the reasoning rationale by critically evaluating the validity of the preliminary judgment, thus achieving a deeper integration of visual information and domain knowledge. We also observed that the generated reflective rationales often contain causal and contrastive conjunctions, suggesting that ReCoT performs structured and interpretable reflective reasoning similar to human metacognitive processes. Although a single reflection round is used for efficiency, it already achieves state-of-the-art performance, indicating that ReCoT effectively captures the essential mechanism of human-like reflective reasoning.

3.3 Answer Refinement

In the final reasoning step, previous methods often rely on retrieval for commonsense knowledge and use image captions for visual grounding, which may result in incomplete or insufficient information and makes it difficult for the small LM to accurately infer the answer.

Since the preliminary answer and CoT rationale generated in the first two steps have extracted the necessary visual and common-sense information and integrated them into the text modality, we are able to infer the final answer accurately without relying on additional external knowledge or visual inputs. Specifically, we construct the following template for the LM in this step:

Question: <Question> \n Candidate: <Candidate>Rationale: <Rationale>

We generate input texts by replacing the placeholder <*Question*> with the current question Q, the placeholder <*Candidate*> with the preliminary answer a_{MLLM} and the placeholder <*Rationale*> with the reflective rationale $R_{a_{MLLM}}$. The constructed input is fed into the LM to generate the final answer a. The specific formula is as follows:

$$a = f_{LM}(Q, a_{MLLM}, R_{a_{MLLM}}), \tag{3}$$

where f_{LM} is a LM, and we employ FLAN-T5 [62] in our implementation to leverage the reasoning capabilities of small models [63]. Formally, the probability of the answer refinement can be expressed as follows:

$$p\left(a \mid Q, a_{MLLM}, R_{a_{MLLM}}\right) = \prod_{i=1}^{N_a} p_{\theta_a}\left(a_i \mid Q, a_{MLLM}, R_{a_{MLLM}}, a_{<i}\right), \tag{4}$$

where θ_a represents the parameters of the LM f_{LM}, and N_a denotes the length of the GT answer a.

This step refines the preliminary answer based on the generated CoT rationale, reflecting the correction mechanism in metacognition. By leveraging the information obtained in the first two steps, we fully endow the small LM with visual and knowledge foresight, as demonstrated in the subsequent visualization experiments.

4 Experiments

4.1 Datasets

In traditional VQA research, several datasets such as GQA [64] have been developed. To evaluate the knowledge-based VQA methods, researchers have constructed several knowledge-based VQA datasets, such as OK-VQA [65], A-OKVQA [66], Encyclopedic-VQA dataset [67] and InfoSeek dataset [68].

OK-VQA dataset. The OK-VQA dataset [65] contains 14,031 images from the 2014 partition of the COCO dataset [36] and 14,055 questions based on these images, of which 9009 are used for training and 5046 for validation. These questions fall into 10 specific knowledge categories or 'other'. Each sample contains one question, one image, and ten human-annotated GT answers.

A-OKVQA dataset. The A-OKVQA dataset [66] contains 23,692 images from the 2017 partition of the COCO dataset [36] and 24,903 questions based on these images, of which 17,056 are used for training, 1145 for validation, and 6702 for testing. The knowledge categories of a randomly sampled subset of 1000 questions in A-OKVQA are approximately 44%, 36%, 18%, and 3% for Visual, Commonsense, Knowledge Base, and Physical, respectively. Each sample contains one question, one image, three human-annotated rationales and ten human-annotated GT answers.

Encyclopedic-VQA dataset. The Encyclopedic-VQA dataset [67] consists of 221K question- answer pairs, primarily sourced from Wikipedia. The questions are categorized into single-hop and two-hop types: single-hop questions require answers from a single Wikipedia page, while two-hop questions necessitate a sequential retrieval process across multiple documents. The dataset is split into 1 M training, 13.6 k validation, and 5.8 k test samples.

InfoSeek dataset. The InfoSeek dataset [68] contains 1.3 M image-question pairs associated with approximately 11 K Wikipedia pages. The dataset is divided into 934 k training, 73 k validation, and 348 k test samples. The validation set includes questions that are not present in the training split and questions linked to unseen entities. Results are reported on the validation set.

4.2 Experimental Setting

Given that LLaVA [33,37] has been extensively pre-trained and instruction-tuned on a variety of VQA benchmarks, it exhibits strong generalization and reasoning capabilities across diverse multimodal queries. This makes it an ideal cognitive agent in our metacognitive reasoning framework, where it is responsible for generating both preliminary answers and reflective rationales. Specifically, we employ LLaVA-1.5 with 13 billion parameters [33,37] as the MLLM in our system. To better demonstrate that our method effectively elicits the MLLM's latent visual and knowledge understanding with respect to the question, we delegate the final answer prediction to a lightweight LM. This design not only highlights the utility of the reflective rationales but also enhances the flexibility and efficiency of the framework by fully fine-tuning the smaller model for the answer refinement stage. For this purpose, we adopt the T5 encoder-decoder architecture [13] under the `base` setting, initialized with FLAN-T5-Base (248 M) [62]. We fine-tune the LM for 20 epochs using the AdamW optimizer, which is widely used for fine-tuning transformer-based models. The learning rate is set to 5×10^{-5} to balance convergence speed and stability. The parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are standard values known to stabilize optimization in transformer models. A weight decay of 0.01 is applied to prevent overfitting while promoting generalization. The batch size is set to 32, a commonly used value in similar tasks to balance computational efficiency and model performance. For the sequence lengths, the maximum input length is set to 512 tokens, which is typical for transformer models, allowing the processing of long sequences. The output sequence length is set to 64 tokens, which is sufficient for the typical answer length in the datasets used. All training and evaluation are conducted on a single NVIDIA A800 GPU with 80 GB of memory, using Python 3.8.12 and PyTorch 1.11.0.

4.3 Experimental Results

Comparative Results on OK-VQA. Table 1 presents a comprehensive comparison between our proposed method and existing state-of-the-art approaches on the OK-VQA dataset. The upper section of the table reports retrieval-based methods that incorporate various forms of external knowledge.

These include methods built on vision-language models, such as ConceptBERT [69], KRISP [5], Visual RR [10], MAVEx [70], and UnifER [54], as well as methods based on pretrained language models, including TRiG [8], RA-VQA [6], and REVEAL [52]. The middle section introduces comparisons with closed-source MLLMs, which represent the frontier of proprietary systems in vision-language reasoning. These include Flamingo [30], Gemini [71], and GPT-4V [72]. The lower section highlights methods that leverage large models, including few-shot prompting approaches using frozen LLMs, such as PICa [58], Prophet [60], and PromptCap [59]; methods based on LLMs augmented with external knowledge, such as REVIVE [14], TwO [15], and SKP [73]; CoT-based reasoning methods such as VCTP [23]; and methods based on large vision-language models enhanced with external signals, exemplified by FLMR [57] and Self-KSel-QAns [74]. The compared results for the closed-source large models are taken from [72], while the results for the other methods are directly reported from the original papers. The 'Pub&Year' column indicates the conference abbreviation and year of publication, the 'R' column denotes whether the method uses CoT rationales for reasoning, and the 'JFT' column indicates whether the method is jointly fine-tuned with frozen large models of 3B parameters or greater.

Table 1: Comparison with state-of-the-art methods on the OK-VQA dataset. The 'Pub&Year' column indicates the conference abbreviation and year of publication. The 'R' column indicates whether the method uses CoT rationales to reason answers. The 'JFT' column indicates whether the method is jointly fine-tuned with frozen large models of 3B or greater. The 'Knowledge Resources' column indicates the main sources of knowledge

Method	Pub&Year	R	JFT	Knowledge Resources*	Accuracy
Retrieval-based methods					
ConceptBERT [69]	EMNLP 2020	×	×	CN	33.66
Krisp [5]	CVPR 2021	×	×	DBpedia + CN + VG + haspartKB	38.35
Visual RR [10]	EMNLP 2021	×	×	Google Search	39.20
MAVEx [70]	AAAI 2022	×	×	CN + WikiPedia + Google Image	40.28
UnifER [54]	ACMMM 2022	×	×	CN	42.13
TRiG [8]	CVPR 2022	×	×	Wikipedia	50.50
RA-VQA [6]	EMNLP 2022	×	×	Google Search	54.48
REVEAL [52]	CVPR 2023	×	×	WIT + CC12M + Wikidata	59.10
Closed-source large models					
Flamingo [30]	–	–	–	–	57.8
Gemini [71]	–	–	–	–	62.27
GPT-4V [72]	–	–	–	–	64.28
Large models-based methods					
PICa-Full [58]	AAAI 2022	×	×	GPT-3	48.00

(Continued)

Table 1 (continued)

Method	Pub&Year	R	JFT	Knowledge Resources*	Accuracy
REVIVE [14]	NeurIPS 2022	✓	×	GPT-3 + Wikidata	56.60
Prophet [60]	CVPR 2023	×	×	GPT-3	61.10
PromptCap [59]	ICCV 2023	×	×	GPT-3	60.40
TwO [15]	ACL 2023	✓	×	GPT-3 + WikiPedia	56.67
FLMR [57]	NeurIPS 2023	×	×	BLIP2 T5-XL + Google Search	62.08
VCTP [23]	AAAI 2024	✓	×	Codex	56.2
SKP [73]	ACL 2024	×	✓	Vicuna + Google Search	63.3
Self-KSel-QAns [74]	EMNLP 2024	×	✓	BLIP2 T5-XL + Google Search	62.8
ReCoT (Ours)		✓	×	BLIP-2 XL	56.76
ReCoT (Ours)		✓	×	LLaVA-1.5	**65.56**†

Note: *CN and VG denote ConceptNet and Visual Genome, respectively. †The best performance is highlighted in bold.

Experimental results show that our method surpasses traditional retrieval-based approaches by 3.48% (65.56 vs. 62.08), outperforms large model-based methods by 2.26% (65.56 vs. 63.30), and further exceeds the best-performing closed-source model by 1.28% (65.56 vs. 64.28). The consistent performance improvements observed in the reflective reasoning stage indicate that ReCoT implicitly performs metacognitive regulation—monitoring, evaluating, and refining its reasoning process—thereby exhibiting metacognitive characteristics in practice. Moreover, compared with methods that jointly fine-tune frozen large models of 3B parameters or greater, our approach still achieves superior performance. Jointly fine-tuning our model with large frozen models (3B parameters or greater) leads to a significant computational burden. Rather than jointly training with large-scale models, ReCoT leverages MLLMs as reasoning agents during inference to execute metacognitive processes. This design enables ReCoT to achieve performance surpassing several baselines that jointly fine-tune models larger than 3B parameters, while significantly reducing computational cost.

Comparative Results on A-OKVQA. In Table 2, we further evaluate the generalization ability of our method on the A-OKVQA dataset, which consists of more challenging knowledge-intensive questions. Our proposed ReCoT framework achieves superior performance compared to recent large model–based approaches. Traditional methods such as ViLBERT [75], LXMERT [11], and KRISP [5] exhibit limited performance due to their shallow visual-textual alignment and restricted reasoning capabilities. Recent large model-based approaches, including PromptCap [59], Prophet [60], and VCTP [23], benefit from pretrained language knowledge but still encounter performance bottlenecks, particularly in complex multimodal reasoning scenarios without metacognitive guidance. The compared results for PromptCap and its earlier methods are from [59], while the more recent works, following PromptCap, are reported from the original papers.

Table 2: Comparison with SOTA methods on the A-OKVQA dataset. The 'Pub&Year' column indicates the conference abbreviation and year of publication. The 'R' column indicates whether the method uses CoT rationales to reason answers. The 'JFT' column indicates whether the method is jointly trained with frozen large models of 3B or greater

Method	Pub&Year	R	JFT	Val	Test
ViLBERT [75]	NeurIPS 2019	×	×	30.6	25.9
LXMERT [11]	EMNLP 2019	×	×	30.7	25.9
KRISP [5]	CVPR 2021	×	×	33.7	27.1
GPV-2 [76]	ECCV 2022	×	×	48.6	40.7
PromptCap [59]	ICCV 2023	×	×	56.3	59.6
Prophet [60]	CVPR 2023	×	×	58.2	55.7
VCTP [23]	AAAI 2024	✓	×	53.2	53.8
SKP [73]	ACL 2024	×	✓	63.8	–
ReCoT (Ours)		✓	×	**69.6**†	**63.7**†

Note: †The best performance is highlighted in bold.

Among these methods, SKP [73] stands out as one of the strongest baselines, achieving 63.8% on the validation set. However, our ReCoT method further improves upon SKP by a significant margin, achieving **69.6%** on the validation set and **63.7%** on the test set. The experimental results confirm that even in tasks requiring complex visual understanding and external knowledge integration, ReCoT effectively incorporates preliminary cognition and reflective reasoning to dynamically refine predictions. This metacognitive mechanism enables our method to robustly handle diverse reasoning demands and achieve state-of-the-art performance on A-OKVQA without relying on heavy joint fine-tuning with large models.

Generalization Evaluation. To demonstrate the model's ability to generalize across different datasets, we conducted experiments on the Encyclopedic-VQA and InfoSeek datasets. These datasets present highly challenging questions that require not only visual recognition but also detailed external knowledge, making them particularly difficult for models that rely solely on visual input. As shown in Table 3, we compare our method, ReCoT, against several advanced LLMs and MLLMs as baselines. The compared results are taken from [77]. Compared to standard large model baselines, ReCoT demonstrates a significant performance improvement. Vanilla models, such as Vicuna-7B and LLaMA, struggle on both the Encyclopedic-VQA and InfoSeek datasets, which require external knowledge and visual input. In contrast, ReCoT, utilizing a metacognitive reasoning approach, achieves higher scores of 18.6 on Encyclopedic-VQA and 11.3 on InfoSeek, showcasing its superior generalization ability and effectiveness in multimodal tasks. When compared to other MLLMs like InstructBLIP and LLaVA-1.5, which also integrate visual and textual input, ReCoT outperforms them on both datasets, underscoring the value of its metacognitive framework in enhancing reasoning and knowledge utilization. This improvement demonstrates ReCoT's effectiveness in handling knowledge-intensive VQA tasks.

Table 3: More evaluation on the encyclopedic-VQA and InfoSeek datasets

Method	LLM	Encyclopedic-VQA	InfoSeek
LLMs			
Vanilla	Vicuna-7B	2.0	0.0
Vanilla	LLaMA-3-8B	17.3	0.0
Vanilla	LLaMA-3.1-8B	16.6	0.0
MLLMs			
InstructBLIP	Flan-T5 XL	12.0	8.1
LLaVA-1.5	Vicuna-7B	16.9	9.5
ReCoT (Ours)	Vicuna-7B	**18.6**†	**11.3**†

Note: †The best performance is highlighted in bold.

Discuss. The superior performance of our proposed ReCoT method mainly stems from its metacognition-inspired reflective reasoning mechanism. Unlike previous approaches that either rely on external knowledge retrieval or direct multimodal fine-tuning, the metacognitive design enables ReCoT to (1) activate and utilize the rich implicit knowledge within MLLMs without heavy joint fine-tuning, (2) generate higher-quality reflective rationales that provide explicit visual-knowledge grounding, and (3) empower smaller models to reason effectively with minimal computational cost. These reflective rationales significantly enhance reasoning quality and stability across all knowledge categories, leading to consistent state-of-the-art performance on above datasets.

4.4 Ablation Study

We conduct a series of ablation experiments to demonstrate the advantages of the proposed ReCoT method.

Comparison with Rationales Generated by LLM. To verify that rationales generated by traditional LLMs struggle to provide comprehensive information for enhancing smaller LMs, we conduct an experiment where the rationales and corresponding answers generated by GPT-3 from KAT [17] are fed, along with the question, into a FLAN-T5 model for fine-tuning. The experimental results are summarized in Table 4.

Table 4: Comparison of ReCoT and GPT-3 in terms of rationale quality

Model	OK-VQA
GPT-3	48
GPT-3 + T5	48.41 (+0.41)
LLaVA-1.5	60.23
ReCoT (Ours)	65.56 (+5.33)

As shown in Table 4, the smaller LM achieves only a marginal improvement of **+0.41%** when utilizing the rationales generated by GPT-3. In contrast, our reflective rationales significantly enhance the performance of LLaVA-1.5, leading to an improvement of **5.33%** over the original results. This demonstrates that simply leveraging traditional LLM-generated rationales provides limited benefits for downstream reasoning, whereas reflective rationales can substantially refine the answer quality.

These results highlight the importance of metacognitive reasoning in rationale generation. Unlike traditional LLM outputs, our reflective rationales dynamically correct and enrich the preliminary reasoning process, providing more targeted and complete guidance for subsequent answer refinement. This validates the effectiveness of the ReCoT framework in improving complex visual reasoning performance.

Comparison with Rationales Generated by LM. To verify that small LMs struggle to generate effective rationales for knowledge-based VQA tasks, we employ the MM-CoT framework [19], which has been commonly used in ScienceQA [22], to generate rationales and infer answers. The model architecture follows the design shown in Fig. 1b. To further explore whether richer inputs can improve rationale quality and answer accuracy, we incorporate different types of captions and additional retrieved knowledge. We consider the following settings and report the results on the A-OKVQA dataset in Table 5.

Table 5: Comparison of ReCoT and MM-CoT in terms of rationale quality

Method	Context	A-OKVQA
MM-CoT [19]	NoCap	26.60
	OFACap	35.97
	OFACap + K	35.91
	PromptCap + K	42.85
ReCoT (Ours)	–	**69.6**†

Note: †The best performance is highlighted in bold.

NoCap. No contextual information is provided.

OFACap. Captions generated by the OFA large model [78], initialized with the official `large-best-clean` checkpoint, are used.

PromptCap. Question-related captions generated by PromptCap [59], designed to capture visual details relevant to the question, are used.

K. Five relevant knowledge triples are retrieved from ConceptNet [79] following the retrieval method in [7].

As shown in Table 5, MM-CoT under the NoCap setting achieves a poor performance of only 26.60%. Providing captions improves performance to 35.97% (OFACap) and 42.85% (PromptCap + K), but even with enhanced captions and additional retrieved knowledge, MM-CoT still falls significantly short compared to our ReCoT, which achieves 69.6%.

These results indicate that simply providing small LMs with more external information is insufficient for generating effective rationales in complex VQA scenarios. In contrast, our method leverages a metacognitive reasoning process, where an MLLM dynamically monitors and refines its initial cognition, resulting in richer and more targeted rationales. This enables ReCoT to substantially outperform methods that rely solely on small LMs for rationale generation.

Category-Wise Effectiveness of Reflective Reasoning. To evaluate the effectiveness of our reflective rationale in refining preliminary answers, we conduct a category-wise analysis on the OK-VQA validation set. Specifically, we report the proportions of correctly refined, incorrectly refined, and accuracy-unchanged cases among modified samples across different question types. The results are illustrated in Fig. 3, where each bar represents the outcome distribution for a particular question category.

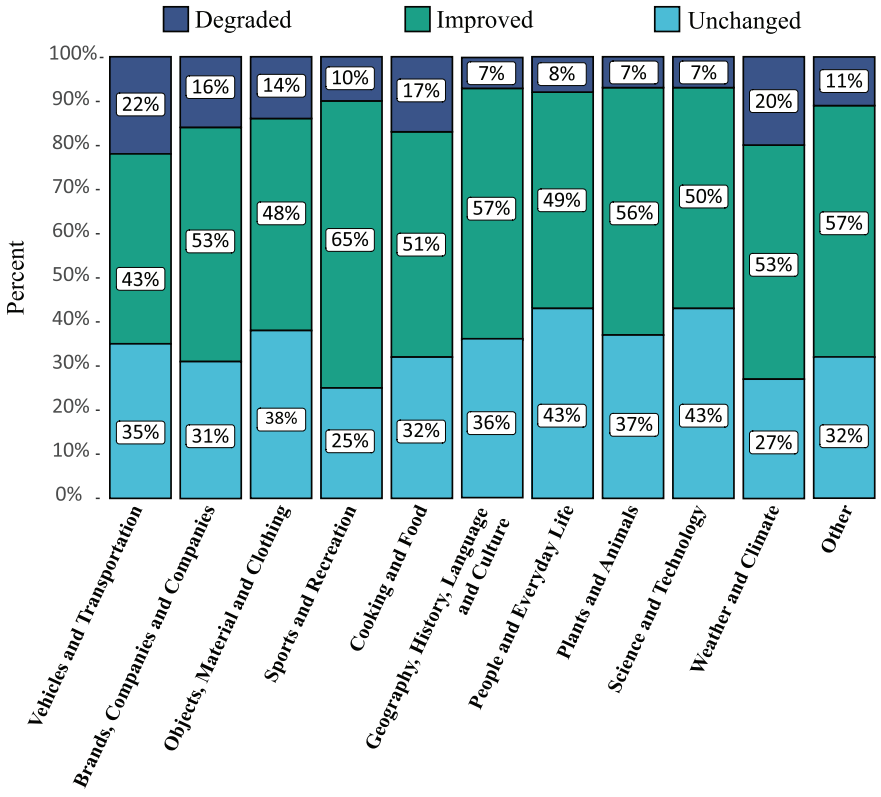

Figure 3: Category-wise distribution of refinement outcomes, showing the proportions of correctly refined, incorrectly refined, and unchanged after reflective reasoning on the OK-VQA validation set

The results demonstrate that our reflective reasoning mechanism consistently improves performance across nearly all categories. Among the modified samples, approximately 50% are correctly refined in each category, indicating the broad applicability of our approach. The category "Sports and Recreation" shows the highest correction rate at 65%, followed by "Geography, History, Language and Culture" (57%) and "Other" (57%). Even in categories such as "Weather and Climate" and "Science and Technology", which often require factual precision, the correction rates reach 53% and 50%, respectively.

In terms of incorrect modifications, the degradation rate remains below 20% in almost all categories, showcasing the stability of the refinement process. The highest degradation occurs in "Vehicles and Transportation" (22%), whereas categories like "Geography, History, Language and Culture", "Plants and Animals" and "Science and Technology" maintain minimal degradation at only 7%.

These results highlight the robustness and generalizability of ReCoT's reflective reasoning process. Its ability to consistently correct errors across diverse semantic domains suggests that the rationale captures key visual and contextual cues overlooked in the initial prediction. Moreover, the low degradation rates indicate that the reflection mechanism introduces minimal risk of overcorrection. This further demonstrates that our metacognitive approach not only enhances reasoning accuracy but also maintains high stability across various knowledge-grounded question types.

Category-Wise Performance Improvement via Reflective Reasoning. To further evaluate the effectiveness of our reflective rationale in refining preliminary answers, we compare the accuracy before and after refinement across different question categories on the OK-VQA validation set. The comparison is visualized in Fig. 4, where the left and right portions of each bar represent the performance of the initial predictions

and their refined results, respectively. The dark red segments indicate the category-wise performance improvements contributed by our reflective reasoning mechanism.

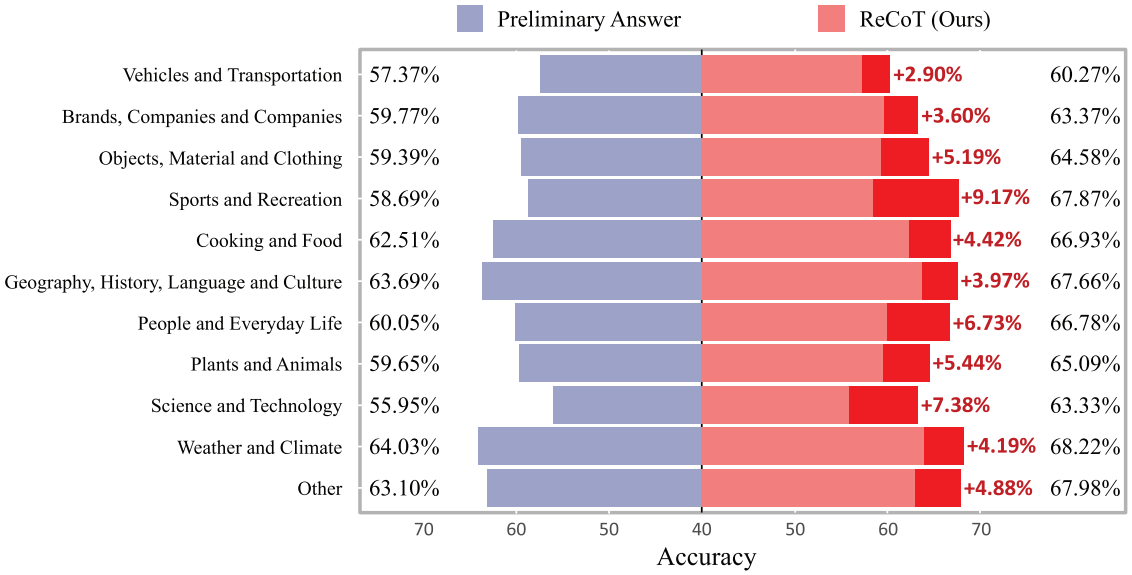

Figure 4: Category-wise accuracy comparison before and after reflective refinement on the OK-VQA validation set. Dark red segments indicate the improvement contributed by our method

As shown in Fig. 4, our method consistently improves accuracy across all categories. The most substantial improvement is observed in the category "Sports and Recreation", with a performance gain of +9.17%, followed by "People and Everyday Life" (+6.73%) and "Science and Technology" (+7.38%). These gains indicate the effectiveness of our approach in both commonsense-rich and knowledge-intensive domains. Even categories with relatively strong initial performance, such as "Weather and Climate" and "Cooking and Food", benefit from improvements of +4.19% and +4.42%, respectively. The smallest improvement is seen in "Vehicles and Transportation" (+2.90%), though performance still increases.

These results demonstrate that our metacognitive reflective reasoning mechanism contributes consistent and meaningful improvements across diverse question types. The substantial gains in both general and knowledge-specific domains suggest that the generated rationales effectively capture overlooked visual or contextual cues, enabling more accurate final predictions. Furthermore, the performance improvements are especially notable in categories where preliminary answers are more likely to suffer from superficial understanding, validating the strength of our refinement strategy.

Effectiveness of Each Component in ReCoT. To assess the contribution of each component within the ReCoT framework, we conduct ablation studies by systematically removing key stages. Specifically, the 'w/o IG' setting directly prompts LLaVA to generate rationales for answer inference without leveraging reflection or refinement of preliminary answers. The 'w/o PA' setting executes both the answer generation and reflective reasoning stages but uses only the rationale for final answer inference, discarding the preliminary answer. Additionally, we assess the impact of incorporating captions and external retrieved knowledge. The experimental results are reported in Table 6.

As shown in Table 6, without reflection on preliminary answers ('w/o IG'), directly prompting LLaVA to generate rationales leads to a substantial performance drop—achieving only 54.82% on OK-VQA, which is **9.36%** lower than our full ReCoT pipeline (65.56%). Removing the influence of preliminary answers

('w/o PA') also causes noticeable degradations in performance across both OK-VQA and A-OKVQA. Furthermore, while incorporating image captions and externally retrieved knowledge results in very slight performance improvements (e.g., 69.98% vs. 69.55% on A-OKVQA), the gains remain marginal compared to the benefits brought by ReCoT's internal reflective reasoning. Although the quality of the captions and retrieved knowledge is substantially lower than that of the reflective rationales generated by our method, they do contain some additional information that is not captured by the reflective rationales. However, this supplementary information only leads to very limited improvements in overall performance.

Table 6: Ablation study of each component in our proposed method

Input	OK-VQA	A-OKVQA
Ours	65.56	69.55
w/o IG	54.82	61.30
w/o PA	64.18	68.19
w/OFACap	65.42	69.64
w/OFACap + K	65.58	69.71
w/PromptCap + K	65.53	69.98

These results further validate the importance of the reflective reasoning mechanism and demonstrate that each stage of our metacognitive reasoning plays an essential and complementary role. Moreover, they indicate that while external knowledge and captions can offer minor enhancements, the internally generated cognition and reflection in ReCoT already provide sufficiently comprehensive information for complex visual reasoning.

Effectiveness of Affording Knowledge and Visual Information. To validate the effectiveness of the proposed ReCoT framework in affording essential information for answer reasoning, we compare it against several retrieval-based state-of-the-art methods under identical settings. All compared methods employ the T5 encoder-decoder architecture to infer answers. For a fair comparison, we initialize our model with the same `T5-large` parameters used by these retrieval-based approaches. The experimental results are summarized in Table 7, where the 'Input' column lists the types of information provided to each method during answer inference.

As shown in Table 7, existing retrieval-based methods, such as TRiG [8], RA-VQA [6], and REVEAL [52], incorporate a wide range of external knowledge sources and visual information, including dense labels, OCR, and image captions. Despite these extensive efforts to gather and integrate external information, their performances remain significantly lower than ours. In particular, ReCoT achieves an accuracy of **65.71%**, substantially outperforming the best retrieval-based baseline (REVEAL, 59.10%).

These results demonstrate that ReCoT can effectively afford the critical knowledge and visual cues needed for complex reasoning tasks by internally generating preliminary answers and reflective rationales, without requiring explicit retrieval or image-to-text conversion. This highlights the strength of our metacognitive reasoning approach in eliciting necessary information within MLLMs compared to traditional retrieval-based pipelines.

Effectiveness across Backbones. To validate the generalizability of our proposed ReCoT framework across different model backbones, we replace the LM with several widely-used variants, including FLAN-Alpaca and UnifiedQA [80]. In addition, we evaluate the performance of our method when using models with visual features or larger parameter sizes. The experimental results are presented in Table 8.

Table 7: Comparison of methods using the same network structure and initialization parameters but different inputs

Method	Input	Accuracy
TRiG [8]	Question + Image caption + Dense labeling + OCR + WikiPedia	50.50
RA-VQA [6]	Question + Image caption + Object name + Object attribute + OCR + Google search	54.48
RIVIVE [14]	Question + Image caption + Regional tag + Visual feature + Wikidata + Candidate + Rationale	56.60
REVEAL [52]	Question + Visual feature + WIT + CC12M + Wikidata + VQA-v2	59.10
FLMR [57]	Question + Image captions + Object name + Object attribute + OCR + Google search	54.85
ReCoT (Ours)	Question + Preliminary answer + Reflective rationale	**65.71**†

Note: †The best performance is highlighted in bold.

Table 8: Using different backbone. VF represents the visual features extracted from images

Backbone	OK-VQA	A-OKVQA
$T5_{base}$	64.85	68.90
FLAN-$Alpaca_{base}$	65.30	69.74
$UnifiedQA_{base}$	64.93	68.47
FLAN-$T5_{base}$	65.56	69.55
FLAN-$T5_{large}$	65.63	69.21
FLAN-$T5_{base}$ + VF	65.16	69.46

As shown in Table 8, our method consistently achieves strong performance across all backbones. Specifically, when replacing the LM with FLAN-$Alpaca_{base}$ or $UnifiedQA_{base}$, the performance remains comparable to that of the original FLAN-$T5_{base}$ model. Moreover, injecting visual features extracted from ViT [81] (FLAN-$T5_{base}$ + VF) or using a larger backbone such as FLAN-$T5_{large}$ brings no significant performance improvement on either OK-VQA or A-OKVQA.

These results demonstrate the robustness and generalizability of ReCoT across diverse LM backbones. The generated preliminary answers and reflective rationales afford sufficient knowledge and visual cues for downstream reasoning, enabling small models to achieve strong performance without the need for additional visual features or larger model capacities. This further highlights the efficiency and flexibility of the ReCoT framework for knowledge-based VQA tasks.

4.5 Qualitative Analysis

To demonstrate how our ReCoT rationales effectively guide answer refinement, we present several representative examples from the OK-VQA validation set in Fig. 5. In the first example, the model's initial error ("Time") arises from **language ambiguity**, as it overlooks the plural context implying multiple zones. Through rational reasoning, it recognizes that the clocks display different time zones, correcting the answer to "Time zone." Similar patterns appear elsewhere: some errors stem from **visual misjudgment** (e.g., misidentifying the ram's habitat or missing the mammal's fur), while others reflect **knowledge gaps** (e.g., incorrect founding year or author). Across examples, the rationales effectively leverage visual and contextual cues to overcome these initial misunderstandings and produce accurate final predictions. These cases demonstrate that ReCoT not only helps recover semantically rich knowledge overlooked in the initial prediction but also enables the model to incorporate grounded visual cues and commonsense knowledge. In addition, we observed that the generated reflective rationales exhibit good language fluency and practical usability, further confirming their interpretability and applicability in multimodal reasoning scenarios. We further analyze several failure cases of ReCoT to better understand its limitations. Some incorrect predictions occur when the model's reflective rationale fails to distinguish between semantically related but factually incorrect concepts or when visual evidence is overemphasized while ignoring external knowledge. These cases suggest that although ReCoT effectively enhances reasoning in most scenarios, its reflective process can still inherit certain biases from the underlying MLLM. This highlights potential directions for improving factual grounding and balancing visual-knowledge integration in future work.

4.6 Difference between ReCoT and Metacognition

AI is closely intertwined with human cognitive science, and understanding human learning and reasoning processes provides valuable inspiration for building more intelligent AI systems. Although ReCoT simulates the core processes of metacognition, a gap remains between ReCoT and a full realization of human metacognitive mechanisms. Specifically, ReCoT currently relies on MLLMs to generate preliminary answers, meaning that its reflective reasoning process is inherently constrained by the capabilities and training knowledge of the underlying models. Furthermore, while metacognitive theory in cognitive science emphasizes continuous and dynamic reflection across multiple iterative stages of the cognitive process, ReCoT primarily focuses on reflecting upon cognitive outcomes without engaging in deeper multi-step iterative refinement. Due to the frozen nature of the MLLM, if an incorrect preliminary answer is produced in the first iteration, it becomes extremely challenging to generate a correct reflective rationale to correct it. As a result, even if multiple iterations of reflection are performed, the probability of successfully refining the answer remains low, while the computational cost increases significantly. Under the current framework, deeper iterative refinement offers limited practical benefits relative to its computational overhead. Nevertheless, we believe that ReCoT provides valuable insights into the construction of intelligent agents guided by principles of human cognition. In future work, we will further explore how to enhance the iterative optimization capability of ReCoT to better align with the continuous reflection mechanisms emphasized in metacognitive theory.

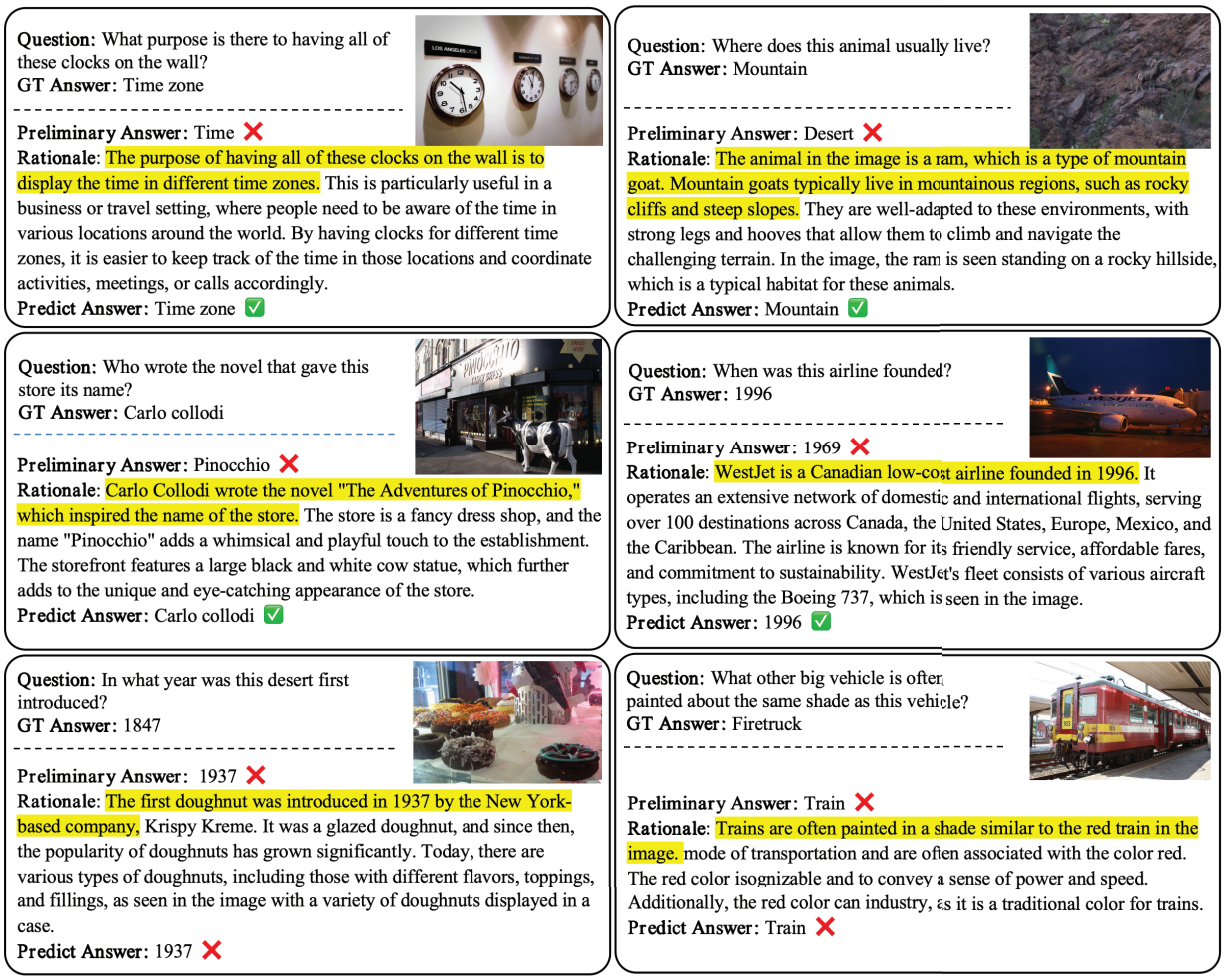

Figure 5: Qualitative examples of ReCoT's reflective reasoning process on the OK-VQA validation set. Each example shows how ReCoT refines incorrect preliminary answers by generating rationales that integrate visual cues and contextual knowledge, resulting in accurate final predictions aligned with GT answers

5 Conclusion

In this paper, we proposed ReCoT, a simple yet effective three-step CoT reasoning framework inspired by metacognition theory. Under the setting where LLaVA serves as the cognitive agent, ReCoT empowers smaller LMs with critical visual and knowledge insights to accurately infer answers. Specifically, our method generates the final answer by reflecting on and refining an initial prediction through the use of reflective rationales. Experimental results demonstrate that ReCoT can effectively stimulate and leverage the internal knowledge and visual understanding embedded within MLLMs, enabling accurate reasoning without the need for explicit retrieval or image-to-text conversion. By harnessing metacognitive reflective processes, ReCoT achieves state-of-the-art performance on knowledge-based VQA benchmarks. Nevertheless, ReCoT currently depends on the capabilities and training knowledge of the underlying MLLMs for both preliminary answer generation and reflective reasoning, inherently limiting the extent of reflection achievable. Enabling continuous and dynamic reflection to progressively refine cognitive outcomes remains a promising direction for future research.

To further bridge the gap between artificial and human metacognition, future work will extend ReCoT into a multi-round reflective reasoning framework with a memory or self-monitoring mechanism, enabling iterative metacognitive cycles for continuous evaluation and refinement. This will bring ReCoT closer to human-like reflective cognition and support deeper, adaptive reasoning. Additionally, applying ReCoT to low-resource languages and cross-cultural knowledge reasoning will improve inclusiveness and generalization. Reflective reasoning can help compensate for limited external knowledge, though aligning it with diverse linguistic and cultural contexts remains challenging. Beyond technical development, ReCoT's metacognitive foundation raises important philosophical and ethical questions on AI self-reflection, cognitive transparency, and alignment with human values. Finally, ReCoT shows potential for cross-domain applications in education, healthcare, and human-AI collaboration, enhancing its theoretical and interdisciplinary impact.

Acknowledgement: We gratefully acknowledge the support of the National Natural Science Foundation of China and the R&D Program of Beijing Municipal Education Commission.

Funding Statement: This research is supported by the National Natural Science Foundation of China (Nos. 62572017, 62441232, 62206007), R&D Program of Beijing Municipal Education Commission (KZ202210005008).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Zhongfan Sun and Yongli Hu; methodology, Zhongfan Sun; software, Kan Guo; validation, Yong Zhang; formal analysis, Kan Guo; investigation, Yongli Hu; resources, Zhongfan Sun; data curation, Zhongfan Sun; writing—original draft preparation, Zhongfan Sun and Kan Guo; writing—review and editing, Yong Zhang; visualization, Yongli Hu; supervision, Yong Zhang; project administration, Yongli Hu; funding acquisition, Yongli Hu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available in the OK-VQA (https://okvqa.allenai.org/), A-OKVQA (https://github.com/allenai/aokvqa), Encyclopedic-VQA (https://github.com/google-research/google-research/tree/master/encyclopedic_vqa), and InfoSeek (https://github.com/open-vision-language/infoseek) repositories (accessed on 26 November 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Shoham Y. Reasoning about change: time and causation from the standpoint of artificial intelligence. Cambridge, MA, USA: MIT Press; 1987.
2. Marcus G. The next decade in AI: four steps towards robust artificial intelligence. arXiv:2002.06177. 2020.
3. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. Behav Brain Sci. 2017;40:e253. doi:10.1017/s0140525x16001837.
4. Bisk Y, Holtzman A, Thomason J, Andreas J, Bengio Y, Chai J, et al. Experience grounds language. arXiv:2004.10151. 2020.
5. Marino K, Chen X, Parikh D, Gupta A, Rohrbach M. Krisp: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2021. p. 14111–21.
6. Lin W, Byrne B. Retrieval augmented visual question answering with outside knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2022. p. 11238–54.

7. Sun Z, Hu Y, Gao Q, Jiang H, Gao J, Sun Y, et al. Breaking the barrier between pre-training and fine-tuning: a hybrid prompting model for knowledge-based VQA. In: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: ACM; 2023. p. 4065–73.
8. Gao F, Ping Q, Thattai G, Reganti A, Wu YN, Natarajan P. Transform-retrieve-generate: natural language-centric outside-knowledge visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 5067–77.
9. Kim W, Son B, Kim I. Vilt: vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 5583–94.
10. Luo M, Zeng Y, Banerjee P, Baral C. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2021. p. 6417–31.
11. Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL; 2019. p. 5100–11.
12. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, et al. Vinvl: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2021. p. 5579–88.
13. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
14. Lin Y, Xie Y, Chen D, Xu Y, Zhu C, Yuan L. REVIVE: regional visual representation matters in knowledge-based visual question answering. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. London, UK: PMLR; 2022. p. 10560–71.
15. Si Q, Mo Y, Lin Z, Ji H, Wang W. Combo of thinking and observing for outside-knowledge VQA. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2023. p. 10959–75.
16. Fu X, Zhang S, Kwon G, Perera P, Zhu H, Zhang Y, et al. Generate then select: open-ended visual question answering guided by world knowledge. In: Findings of the association for computational linguistics: ACL 2023. Stroudsburg, PA, USA: ACL; 2023. p. 2333–46.
17. Gui L, Wang B, Huang Q, Hauptmann AG, Bisk Y, Gao J. KAT: a knowledge augmented transformer for vision-and-language. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL; 2022. p. 956–68.
18. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Proc Syst. 2020;33:1877–901.
19. Zhang Z, Zhang A, Li M, Zhao H, Karypis G, Smola A. Multimodal chain-of-thought reasoning in language models. arXiv:2302.00923. 2023.
20. Wang L, Hu Y, He J, Xu X, Liu N, Liu H, et al. T-sciq: teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. Palo Alto, CA, USA: AAAI Press; 2024. p. 19162–70.
21. Li Q, Sun H, Xiao F, Wang Y, Gao X, Bhanu B. PS-CoT-adapter: adapting plan-and-solve chain-of-thought for ScienceQA. Sci China Inf Sci. 2025;68(1):119101. doi:10.1007/s11432-024-4211-9.
22. Lu P, Mishra S, Xia T, Qiu L, Chang KW, Zhu SC, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. Adv Neural Inf Proc Syst. 2022;35:2507–21.
23. Chen Z, Zhou Q, Shen Y, Hong Y, Sun Z, Gutfreund D, et al. Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. Palo Alto, CA, USA: AAAI Press; 2024. p. 1254–62.
24. Hong R, Lang J, Xu J, Cheng Z, Zhong T, Zhou F. Following clues, approaching the truth: explainable micro-video rumor detection via chain-of-thought reasoning. In: Proceedings of the ACM on Web Conference 2025. New York, NY, USA: ACM; 2025. p. 4684–98.

25. Flavell JH. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. Am Psychol. 1979;34(10):906–11. doi:10.1037/0003-066x.34.10.906.
26. Brown A. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In: Metacogniton, motivation, and understanding. Mahwah, NJ, USA: Lawrence Erlbaum Associates; 1987.
27. Nelson TO. Metamemory: a theoretical framework and new findings. In: Psychology of learning and motivation. Vol. 26. Amsterdam, The Netherlands: Elsevier; 1990. p. 125–73.
28. Evans JSB, Stanovich KE. Dual-process theories of higher cognition: advancing the debate. Perspect Psychol Sci. 2013;8(3):223–41. doi:10.1177/1745691612460685.
29. Sweller J. Cognitive load theory. In: Psychology of learning and motivation. Vol. 55. Amsterdam, The Netherlands: Elsevier; 2011. p. 37–76.
30. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. Adv Neural Inf Proc Syst. 2022;35:23716–36.
31. Li J, Li D, Savarese S, Hoi S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. London, UK: PMLR; 2023. p. 19730–42.
32. Ye Q, Xu H, Xu G, Ye J, Yan M, Zhou Y, et al. mPlug-owl: modularization empowers large language models with multimodality. arXiv:2304.14178. 2023.
33. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 26296–306.
34. Changpinyo S, Sharma P, Ding N, Soricut R. Conceptual 12M: pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2021. p. 3558–68.
35. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: an 800GB dataset of diverse text for language modeling. arXiv:2101.00027. 2020.
36. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference; 2014 sep 6–12; Zurich, Switzerland. Cham, Switzerland: Springer; 2014. p. 740–55.
37. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inf Proc Syst. 2023;36:34892–916.
38. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. Uniter: universal image-text representation learning. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2020. p. 104–20.
39. Li P, Gao Z, Zhang B, Yuan T, Wu Y, Harandi M, et al. Fire: a dataset for feedback integration and refinement evaluation of multimodal models. Adv Neural Inf Proc Syst. 2025;37:101618–40.
40. Chen J, Hei X, Xue Y, Wei Y, Xie J, Cai Y, et al. Learning to correction: explainable feedback generation for visual commonsense reasoning distractor. In: Proceedings of the 32nd ACM International Conference on Multimedia. New York, NY, USA: ACM; 2024. p. 8209–18.
41. Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. arXiv:2210.03493. 2022.
42. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2022. p. 22199–213.
43. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2022. p. 24824–37.
44. Nye M, Andreassen AJ, Gur-Ari G, Michalewski H, Austin J, Bieber D, et al. Show your work: scratchpads for intermediate computation with language models. arXiv:2112.00114. 2021.
45. Zheng G, Yang B, Tang J, Zhou HY, Yang S. Ddcot: duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Adv Neural Inf Proc Syst. 2023;36:5168–91.
46. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. React: synergizing reasoning and acting in language models. arXiv:2210.03629. 2023.

47. Mondal D, Modi S, Panda S, Singh R, Rao GS. Kam-cot: knowledge augmented multimodal chain-of-thoughts reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. Palo Alto, CA, USA: AAAI Press; 2024. p. 18798–806.
48. Chowdhury S, Soni B. Beyond words: ESC-Net revolutionizes VQA by elevating visual features and defying language priors. Comput Intell. 2024;40(6):e70010. doi:10.1111/coin.70010.
49. Chowdhury S, Soni B. R-VQA: a robust visual question answering model. Knowl Based Syst. 2025;309(28):112827. doi:10.1016/j.knosys.2024.112827.
50. Xie J, Fang W, Cai Y, Huang Q, Li Q. Knowledge-based visual question generation. IEEE Trans Circ Syst For Video Technol. 2022;32(11):7547–58. doi:10.1109/tcsvt.2022.3189242.
51. Wen Z, Peng Y. Multi-level knowledge injecting for visual commonsense reasoning. IEEE Trans Circ Syst For Video Technol. 2020;31(3):1042–54. doi:10.1109/tcsvt.2020.2991866.
52. Hu Z, Iscen A, Sun C, Wang Z, Chang KW, Sun Y, et al. Reveal: retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 23369–79.
53. Ding Y, Yu J, Liu B, Hu Y, Cui M, Wu Q. Mukea: multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 5089–98.
54. Guo Y, Nie L, Wong Y, Liu Y, Cheng Z, Kankanhalli M. A unified end-to-end retriever-reader framework for knowledge-based VQA. In: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2022. p. 2061–9.
55. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2020. p. 6769–81.
56. Luo L, Lai H, Pan Y, Yin J. Efficient multimodal selection for retrieval in knowledge-based visual question answering. IEEE Trans Circ Syst Video Technol. 2025;35(6):5195–207. doi:10.1109/tcsvt.2025.3527032.
57. Lin W, Chen J, Mei J, Coca A, Byrne B. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. Adv Neural Inf Proc Syst. 2024;36:1–21.
58. Yang Z, Gan Z, Wang J, Hu X, Lu Y, Liu Z, et al. An empirical study of GPT-3 for few-shot knowledge-based VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. Palo Alto, CA, USA: AAAI Press; 2022. p. 3081–9.
59. Hu Y, Hua H, Yang Z, Shi W, Smith NA, Luo J. Promptcap: prompt-guided task-aware image captioning. arXiv:2211.09699. 2022.
60. Shao Z, Yu Z, Wang M, Yu J. Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 14974–83.
61. Wang H, Ge W. Q&A prompts: discovering rich visual clues through mining question-answer prompts for VQA requiring diverse world knowledge. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2024. p. 274–92.
62. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res. 2024;25(70):1–53.
63. Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot task generalization. arXiv:2110.08207. 2021.
64. Hudson DA, Manning CD. GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2019. p. 6700–9.
65. Marino K, Rastegari M, Farhadi A, Mottaghi R. OK-VQA: a visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2019. p. 3195–204.

66. Schwenk D, Khandelwal A, Clark C, Marino K, Mottaghi R. A-OKVQA: a benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2022. p. 146–62.
67. Mensink T, Uijlings J, Castrejon L, Goel A, Cadar F, Zhou H, et al. Encyclopedic VQA: visual questions about detailed properties of fine-grained categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 3113–24.
68. Chen Y, Hu H, Luan Y, Sun H, Changpinyo S, Ritter A, et al. Can pre-trained vision and language models answer visual information-seeking questions? arXiv:2302.11713. 2023.
69. Gardères F, Ziaeefard M, Abeloos B, Lecue F. Conceptbert: concept-aware representation for visual question answering. In: Findings of the association for computational linguistics: EMNLP 2020. Stroudsburg, PA, USA: ACL; 2020. p. 489–98.
70. Wu J, Lu J, Sabharwal A, Mottaghi R. Multi-modal answer validation for knowledge-based VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. Palo Alto, CA, USA: AAAI Press; 2022. p. 2712–21.
71. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv:2312.11805. 2023.
72. Li Y, Wang L, Hu B, Chen X, Zhong W, Lyu C, et al. A comprehensive evaluation of GPT-4V on knowledge-intensive visual question answering. arXiv:2311.07536. 2023.
73. Wang Q, Ji R, Peng T, Wu W, Li Z, Liu J. Soft knowledge prompt: help external knowledge become a better teacher to instruct LLM in knowledge-based VQA. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2024. p. 6132–43.
74. Hao D, Wang Q, Guo L, Jiang J, Liu J. Self-bootstrapped visual-language model for knowledge selection and question answering. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2024. p. 1857–68.
75. Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Palo Alto, CA, USA: AAAI Press; 2019. p. 13–23.
76. Kamath A, Clark C, Gupta T, Kolve E, Hoiem D, Kembhavi A. Webly supervised concept expansion for general purpose vision models. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2022. p. 662–81.
77. Cocchi F, Moratelli N, Cornia M, Baraldi L, Cucchiara R. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. Piscataway, NJ, USA: IEEE; 2025. p. 9199–209.
78. Wang P, Yang A, Men R, Lin J, Bai S, Li Z, et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. London, UK: PMLR; 2022. p. 23318–40.
79. Liu H, Singh P. ConceptNet—a practical commonsense reasoning tool-kit. BT Technol J. 2004;22(4):211–26.
80. Khashabi D, Min S, Khot T, Sabharwal A, Tafjord O, Clark P, et al. UNIFIEDQA: crossing format boundaries with a single QA system. In: Findings of the association for computational linguistics: EMNLP 2020. Stroudsburg, PA, USA: ACL; 2020. p. 1896–907.
81. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2021.