II-MMR: Identifying and Improving Multi-modal Multi-hop Reasoning in Visual Question Answering

Jihyung Kil[1]* **Farideh Tavazoee**[2] **Dongyeop Kang**[3] **Joo-Kyung Kim**[2]

[1]The Ohio State University [2]Amazon AGI [3]University of Minnesota

kil.5@osu.edu {fayt,jookyk}@amazon.com dongyeop@umn.edu

Abstract

Visual Question Answering (VQA) often involves diverse reasoning scenarios across Vision and Language (V&L). Most prior VQA studies, however, have merely focused on assessing the model's overall accuracy without evaluating it on different reasoning cases. Furthermore, some recent works observe that conventional Chain-of-Thought (CoT) prompting fails to generate effective reasoning for VQA, especially for complex scenarios requiring multi-hop reasoning. In this paper, we propose **II-MMR**, a novel idea to **i**dentify and **i**mprove **m**ulti-modal **m**ulti-hop **r**easoning in VQA. In specific, II-MMR takes a VQA question with an image and finds a reasoning path to reach its answer using two novel language promptings: (i) answer prediction-guided CoT prompt, or (ii) knowledge triplet-guided prompt. II-MMR then analyzes this path to identify different reasoning cases in current VQA benchmarks by estimating **how many hops** and **what types** (i.e., visual or beyond-visual) of reasoning are required to answer the question. On popular benchmarks including GQA and A-OKVQA, II-MMR observes that most of their VQA questions are easy to answer, simply demanding "single-hop" reasoning, whereas only a few questions require "multi-hop" reasoning. Moreover, while recent V&L models struggle with such complex multi-hop reasoning questions even using the traditional CoT method, II-MMR shows its effectiveness across all reasoning cases in both zero-shot and fine-tuning settings.[1]

1 Introduction

Reasoning is a key aspect of highly intelligent systems. Visual question answering (VQA) (Goyal et al., 2017; Schwenk et al., 2022; Hudson and Manning, 2019) enables us to measure such reasoning ability as it contains different reasoning scenarios

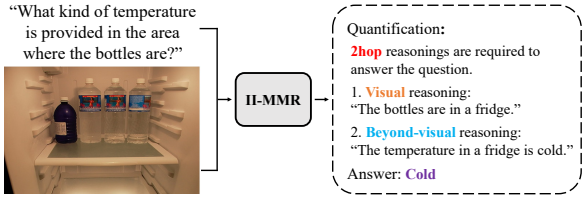


Figure 1: **Overview of II-MMR.** Our II-MMR automatically identifies different reasoning cases in VQA benchmarks by measuring how many and what types (visual or beyond-visual) of reasoning are required to solve a VQA question. The identified reasoning process in II-MMR also helps make a correct prediction (Cold), while the simple Chain-of-Thought (CoT) method (Kojima et al., 2022) fails to answer.

in the benchmark. For instance, a VQA question *"What color is the banana?"* requires one-hop (one-step) reasoning to be answered, which is to identify the color of the banana in the image. In contrast, the other question, *"Which American president is associated with the stuffed animal seen here?"* asks two-hop reasoning: (i) visually detecting this animal as "Teddy bear", and (ii) knowing that the "Teddy bear" is related to President "Roosevelt" (i.e., commonsense reasoning).

Despite different reasoning approaches being required for different questions, most prior VQA studies (Tan and Bansal, 2019; Chen et al., 2023; Wang et al., 2022) have solely focused on the model's *overall* accuracy, neglecting to evaluate its reasoning capabilities. While a few works (Li et al., 2018; Wu and Mooney, 2019) attempt to interpret its reasoning process, they often rely on human explanations, which are challenging to collect sufficiently. Moreover, most VQA benchmarks (Goyal et al., 2017; Schwenk et al., 2022) do not provide detailed information on reasoning, including how many and what types of reasoning are required to answer the question. These limitations hinder the extensive and in-depth understanding of the model's reasoning abilities. Recently, Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022), which elicits complex multi-

* Work was partially done during the Amazon internship.
[1]https://github.com/heendung/II-MMR

Findings of the Association for Computational Linguistics ACL 2024, pages 10698–10709

hop reasoning via step-by-step instructions, has demonstrated remarkable performance across various NLP domains, including arithmetic and logical reasoning (Cobbe et al., 2021; Srivastava et al., 2023). However, some recent studies (Awal et al., 2023) point out that this CoT reasoning may be ineffective for VQA tasks due to (i) the model's limited capabilities to ground visual objects in rationale generation and (ii) its proneness to hallucinate non-existent objects in the image.

Proposal. In this paper, we propose **II-MMR**, a novel method which automatically **i**dentifies and **i**mproves **m**ulti-modal **m**ulti-hop **r**easoning for VQA tasks. Given a VQA question with its relevant image, our II-MMR first identifies a reasoning path to reach its answer using two novel prompting strategies: (i) answer prediction-guided CoT (II-MMR$_{\text{APCOT}}$), or (ii) knowledge triplet-guided prompt (II-MMR$_{\text{KTPROMPT}}$). Concretely, II-MMR$_{\text{APCOT}}$ first prompts a V&L model to directly predict an answer for the question and then generates an answer-related path by incorporating this prediction into the CoT prompt, guiding reasoning towards the answer. Besides, II-MMR$_{\text{KTPROMPT}}$ asks an LLM to extract knowledge triplets from question and answer (QA) and treats the sequence of these triplets as the answer reasoning path. In short, II-MMR utilizes additional cues, either through answer prediction or QA-related knowledge triplets, to find the correct reasoning path.

Effectiveness of II-MMR**.** II-MMR analyzes this reasoning path to identify different reasoning cases in current VQA benchmarks by measuring the number of reasoning steps and the types of reasoning, such as visual or beyond-visual (e.g., commonsense, knowledge base (Schwenk et al., 2022)), required to answer the question (Figure 1). During our prompting process, the intermediate reasoning steps and the alignment of question keywords with visual objects determine the number and types of reasoning.

II-MMR finds two shortcomings of GQA (Hudson and Manning, 2019) and A-OKVQA (Schwenk et al., 2022) benchmarks: (i) a scarcity of multi-hop reasoning questions and (ii) an overestimation of VQA performance due to the high model accuracy on simple one-hop reasoning questions. Concretely, while the current well-known V&L model (e.g., BLIP-2 Li et al. (2023)) excels in such one-hop reasoning questions, it struggles in complex multi-hop scenarios, even using the standard CoT reasoning (Kojima et al., 2022). On the other hand, II-MMR with the proposed language promptings shows notable performance across all reasoning scenarios, including multi-hop cases in both zero-shot and fine-tuning settings.

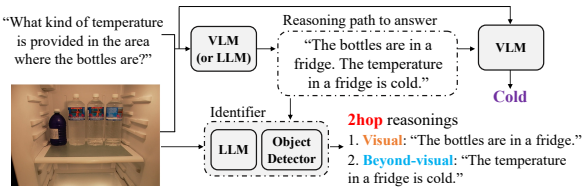


Figure 2: **Pipeline of II-MMR.** Given a VQA question with its image, II-MMR first generates a reasoning path to the answer either using the V&L model (VLM) or the LLM. We then utilize this path to identify different reasoning cases in VQA benchmarks by estimating the number and types (visual or beyond-visual) of reasoning required for the question. Finally, II-MMR feeds the reasoning path, along with the question and the image, into VLM to predict the answer.

In short, our II-MMR suggests that identifying (or breaking down) reasoning helps a better understanding of the internal reasoning process and improves the reasoning performance in multi-hop scenarios. Moreover, we believe our II-MMR could be used to create a more complex and practical multi-hop VQA dataset for future work.

Our main contributions are three-folded:

- We introduce II-MMR to identify and improve the multi-hop reasoning for VQA tasks (Figure 1).
- II-MMR finds that current VQA benchmarks have some flaws, including the shortage of multi-hop reasoning questions and the inflated results due to simple reasoning cases.
- II-MMR shows its effectiveness in all reasoning cases, including multi-hop reasoning ones in both zero-shot and fine-tuning settings.

2 Proposed Approach: II-MMR

Figure 2 provides a pipeline of II-MMR. In what follows, we first describe our two novel promptings, II-MMR$_{\text{APCOT}}$ and II-MMR$_{\text{KTPROMPT}}$, to find a reasoning path leading to the answer (§2.1). Then, we describe how to identify different reasoning cases in current VQA benchmarks (§2.2) using the detected reasoning paths in §2.1. Finally, we discuss how to utilize the reasoning paths to further improve the reasoning performance in zero-shot and fine-tuning stages (§2.3).

2.1 Finding a reasoning path to the answer

2.1.1 Preliminary Analysis

One approach to identifying the reasoning path involves utilizing rationales generated by the large

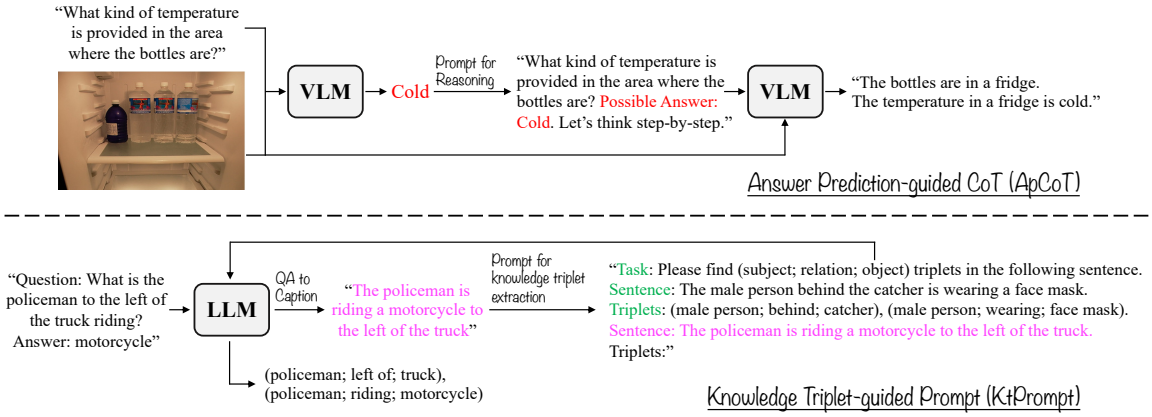


Figure 3: **The language promptings of our II-MMR. Top**: II-MMR$_{\text{ApCoT}}$ first asks the VLM to predict an answer for a VQA question. It then integrates its prediction into the CoT prompt to generate an answer-related rationale, a sequence of reasoning sentences. **Bottom**: II-MMR$_{\text{KtPrompt}}$ initially instructs the LLM to convert the question and answer (QA) to the caption. Then, II-MMR$_{\text{KtPrompt}}$ inputs a prompt (with task, in-context example, and target caption) to the LLM to extract knowledge triplets from QA. We treat the sequence of sentences (or knowledge triplets) as the reasoning path to reach the answer.

Model	A-OKVQA
BLIP-2	46.05
BLIP-2+CoT	36.06

Table 1: **Weakness of traditional CoT prompting.** The zero-shot performance of BLIP-2 on A-OKVQA becomes worse when the conventional CoT reasoning is applied.

language models (LLMs) through CoT prompting. However, the conventional CoT prompting (Kojima et al., 2022) may not be as effective for VQA tasks as for NLP tasks. Indeed, we have observed that BLIP-2 (Li et al., 2023), one of the prominent V&L models, performs significantly worse on A-OKVQA and GQA when employing CoT reasoning, compared to standard prompting (e.g., A-OKVQA: 36.06% vs. 46.05% in Table 1, GQA: 39.08% vs. 44.63% in Table 6). Throughout our comprehensive error analyses, we find that V&L models with CoT fail due to incorrectly or irrelevantly generated rationales. Thus, the rationales generated by the conventional CoT may not be suitable for the reasoning path to the answer.

2.1.2 Answer prediction-guided CoT (ApCoT)

To find a better reasoning path, we introduce an answer prediction-guided CoT (II-MMR$_{\text{ApCoT}}$), which assists the model in generating more answer-related rationales by providing its initial predictions as input context. Concretely, II-MMR$_{\text{ApCoT}}$ starts by prompting a V&L model to directly predict an answer for the VQA question. It then incorporates this predicted answer into the CoT prompt to generate a rationale (Top in Figure 3). We empirically see that the context of the initial prediction leads the model to focus on a topic relevant to the answer and generates a more answer-related rationale. We treat this rationale (concretely, a sequence of reasoning sentences) as a reasoning path to the answer.

Compared to the traditional CoT prompting, our II-MMR$_{\text{ApCoT}}$ notably improves the model answer accuracy, indirectly demonstrating the high quality of our rationales (See §4.3 and §4.4 for more details). Additionally, we conducted a human study with 300 randomly selected VQA samples with our generated rationales and showed the high quality of our rationales. Please see §4.2 for more details.

2.1.3 Knowledge Triplet-guided Prompt (KtPrompt)

In NLP, a knowledge triplet can be viewed as a one-hop (one-step) reasoning. For example, a question, *"Which team does the player named 2015 Diamond Head Classic's MVP play for?"* requires two reasoning steps, *"Buddy Hield is MVP for Diamond Head Classic"* and *"Buddy Hield plays for Sacramento Kings"*. This two-step reasoning can be naturally formed into two knowledge triplets, *(Buddy Hield, MVP, Diamond Head Classic)* and *(Buddy Hield, PlayFor, Sacramento Kings)*.

Built upon this insight, II-MMR$_{\text{KtPrompt}}$ aims to extract knowledge triplets from the question (with the answer) to identify a reasoning path leading to the answer (Bottom in Figure 3). Concretely, II-MMR$_{\text{KtPrompt}}$ first utilizes an LLM (Llama-2-70b Touvron et al. (2023)) to convert the combination of question and answer into a natural caption. Next, we construct an in-context prompt to instruct the LLM to extract knowledge triplets from the

provided caption. However, the LLM may generate noisy knowledge triplets. For instance, triplets may lack some components (e.g., subject, relation, or/and object) or contain trivial words like stopwords (e.g., "the"), which are not typically considered as components. We filter out such noisy samples and obtain a clean set of knowledge triplets. For every VQA question, we treat the sequence of knowledge triplets as a reasoning path to its answer.

2.2 Analyzing a reasoning path

After obtaining the answer reasoning path for each VQA question, we analyze this path to identify different reasoning cases in current VQA benchmarks by automatically measuring the number and types of reasoning required to answer the question. Specifically, we count one reasoning sentence (or one knowledge triplet) in the path as one reasoning step. Besides, the reasoning sentence is categorized into "visual" or "beyond-visual" (Figure 4). We first task an LLM (Llama-2-70b (Touvron et al., 2023)) with extracting keywords from a reasoning sentence. Concurrently, we input the image into GLIP (Li et al., 2022), a phrase-region grounded object detector, to identify objects in the image. We then check how many keywords in the sentence match the visual objects. If *all* keywords match visual objects, the sentence is classified as "visual" reasoning, as it only contains knowledge about the image. If *not all* keywords match, we categorize it as "beyond-visual" reasoning since it involves additional knowledge (e.g., commonsense) beyond the visual information.

2.3 Model performance on the reasoning cases

We investigate whether our II-MMR$_{\text{APCoT}}$ and II-MMR$_{\text{KTPROMPT}}$ effectively improve the model answer accuracy in all the reasoning cases identified in VQA benchmarks (§2.2). Concretely, after obtaining the rationale (or knowledge triplet) through our methods, we prepend them to an answer-triggering prompt (e.g., "Therefore, short answer:"). This combined prompt (with the image and the question) is then fed into the V&L model to make a prediction for the question in each reasoning case. We explore two settings: (i) zero-shot, where the pre-trained model is not further trained with the downstream VQA benchmarks, and (ii) fine-tuning, which involves utilizing the downstream VQA training data to train the model.

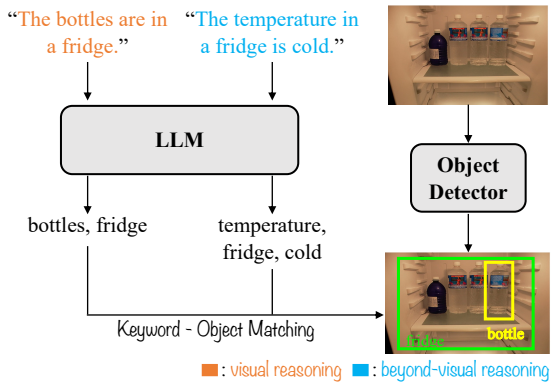


Figure 4: **Analyzing the reasoning types.** The LLM extracts keywords (e.g., "bottle", "temperature") from each reasoning sentence. Meanwhile, the object detector identifies objects (e.g., "fridge", "bottle") in the image. We then check if all keywords match visual objects and decide the reasoning type (visual or beyond-visual) of each sentence in the rationale.

3 Experimental Setup

VQA benchmarks. There exists a variety of VQA benchmarks (Antol et al., 2015; Goyal et al., 2017; Marino et al., 2019; Agrawal et al., 2018; Schwenk et al., 2022; Hudson and Manning, 2019). We explore them in detail and select two VQA benchmarks GQA and A-OKVQA, which most fit with our aim of analyzing different reasoning scenarios. Concretely, GQA is designed to provide compositional reasoning questions over images. A-OKVQA focuses on knowledge-based VQA, requiring knowledge outside images, including commonsense and knowledge base. Due to their design purposes, GQA and A-OKVQA contain multi-hop reasoning questions, useful for analyzing different reasoning cases and understanding the model's reasoning capabilities in various aspects. Please see the appendix for more details about the datasets.

Baselines. We mainly conduct our studies with BLIP-2 (Li et al., 2023), one of the prominent V&L models equipped with strong zero-shot CoT capabilities. We provide two baselines: BLIP-2 and BLIP-2+CoT, which are models without/with traditional CoT reasoning, respectively.

Evaluation metric. We follow the same evaluation metrics used in GQA (Hudson and Manning, 2019) and A-OKVQA (Schwenk et al., 2022). GQA uses the standard accuracy while the A-OKVQA accuracy is based on the average score over nine subsets of the ground-truth ten answers, where each score is: $min(\frac{\#answer\ occurrences}{3}, 1)$.

Training details. We follow the official configu-

rations and implementations of BLIP-2[2]. Specifically, we select the largest BLIP-2 model, which its vision encoder is ViT-g (Fang et al., 2023) and its LLM is FlanT5$_{XXL}$ (Chung et al., 2024). During fine-tuning, the parameters of BLIP-2 are optimized with the language modeling loss on the downstream VQA training data with a batch size of 16 and a learning rate of 1e-5 for ten epochs. We utilize eight A100 48GB GPUs for both training and inference. See the appendix for more details.

Identifying different reasoning cases. We note that the number of reasoning steps in GQA is measured based on the "ground-truth" reasoning path derived from its scene graph (Krishna et al., 2017). GQA questions were generated using scene graph information, such as object relations or attributes. Thus, we leverage this information to obtain the ground-truth path and identify the reasoning cases in GQA, rather than using our generated reasoning path. For A-OKVQA, as no ground-truth paths (or scene graph) exist, we rely on the reasoning paths generated by our II-MMR$_{\text{APCOT}}$. To accurately measure the number of steps, instead of incorporating the model's prediction into the CoT prompt, we include the ground truth answer to generate the answer reasoning path.

4 Experimenetal Results

Aim of our experiments. Our main experimental goals are two-folded: (i) to provide comprehensive statistics on different reasoning cases in current VQA benchmarks (§4.1-§4.2) and (ii) to conduct a thorough analysis of the performance of our II-MMR in these reasoning cases (§4.3-§4.6).

4.1 Analysis of reasoning in VQA benchmarks

Table 2 (Top row) shows the distribution of reasoning steps that our II-MMR identified in GQA and A-OKVQA benchmarks. We note that most GQA questions are simple: requiring direct reasoning (0-hop) (48.74%), involving only the detection of an object in the image (e.g., "Is this a truck?"), or 1-hop reasoning (47.53%). Similarly, 1-hop reasoning questions (69.03%) dominate in A-OKVQA while both benchmarks lack multi-hop reasoning questions (2-hop in GQA: 3.73%, 2 or more-hops in A-OKVQA: 30.97%). These indicate that those VQA benchmarks are biased to evaluate the model's reasoning capabilities in *simple* cases.

[2] https://github.com/salesforce/LAVIS/tree/main

Metric	GQA				A-OKVQA		
	0-hop	1-hop	2-hop	All	1-hop	≥2-hop	All
Hop Distribution	48.74	47.53	3.73	100	69.03	30.97	100
BLIP-2 Accuracy	49.62	42.41	7.66	44.63	46.70	46.54	46.05

Table 2: **Hop distribution and model accuracy on GQA and A-OKVQA**. Simple questions (0/1-hop) dominate, while only a few require multi-hop reasoning (e.g., 2-hop). For zero-shot GQA, the overall accuracy is highly biased to the accuracy of "simple" questions. In contrast, the model suffers in complex questions requiring multi-hop reasoning.

Reasoning Type	A-OKVQA	
	1-hop	2-hop
Visual	37.14	36.69
Beyond-visual	62.85	63.31

Table 3: **Distribution of reasning type on A-OKVQA.** Most questions require knowledge (e.g., commonsense) beyond visual information, aligned with the purpose of this task.

We further analyze the types of reasoning required in A-OKVQA (cf. §2.2). As shown in Table 3, the distribution of reasoning type is skewed toward "beyond-visual" reasoning, suggesting that many questions require knowledge beyond the image to be answered. This finding aligns with the objective of the A-OKVQA task, which assesses the V&L model's knowledge outside the image, such as commonsense or knowledge bases.

In addition, some zero-shot VQA performances are overestimated by the high accuracies on simple questions (Bottom row in Table 2). For instance, BLIP-2 severely suffers on multi-hop reasoning (e.g., 7.66% on 2-hop). However, since direct/1-hop reasoning samples dominate in GQA (48.74%/47.53%) and the model accuracy on these samples is high (49.62%/42.41%), the overall accuracy remains relatively high (44.63%). This suggests that overall accuracy is inflated by the accuracy of simple questions, and thus, relying solely on the overall accuracy may be insufficient to accurately evaluate the model's reasoning abilities.

4.2 Accuracy of predicting hops and reasoning path

As mentioned in §3, for GQA, we are able to obtain the ground-truth reasoning path from its scene graph. Thus, we can evaluate the quality of our reasoning path against the ground-truth one. Table 4 shows that our II-MMR$_{\text{KTPROMPT}}$ is capable of estimating different numbers of reason-

Model	GQA			
	0-hop	1-hop	2-hop	All
II-MMR$_{\text{KTPROMPT}}$	90.55	87.26	84.04	88.74

Table 4: **Hop prediction on GQA.** Our II-MMR$_{\text{KTPROMPT}}$ can estimate the number of reasoning steps required to answer questions over all different reasoning cases.

Model	GQA	
	Strict Matching	Partial Matching
II-MMR$_{\text{KTPROMPT}}$	91.65	94.44

Table 5: **Accuracy of our reasoning path on GQA.** Our reasoning path generated by II-MMR$_{\text{KTPROMPT}}$ highly matches the ground-truth reasoning path, showing its high quality.

ing steps required for questions. For instance, II-MMR$_{\text{KTPROMPT}}$ correctly predicts the number of hops for 88.74% of the total samples in GQA. Moreover, we evaluate the correctness of our reasoning path against the ground-truth path using two matching metrics, Strict and Partial. The former ensures all components in each triplet match between our and ground-truth paths, while the latter is a relaxed version, checking if two of the components match. As depicted in Table 5, our reasoning path is highly consistent with the ground-truth path, demonstrating the benefit of II-MMR$_{\text{KTPROMPT}}$.

We note that as the ground-truth GQA path is formatted as a sequence of knowledge triplets, we only compare it with II-MMR$_{\text{KTPROMPT}}$, which shares the same format, rather than II-MMR$_{\text{APCOT}}$, which utilizes a different format (i.e., the sequence of rationale sentences). For II-MMR$_{\text{APCOT}}$, we instead conduct a human study to evaluate the quality of its generated rationales. We provide each of the three annotators with 100 A-OKVQA samples along with their rationales and focus on two aspects: (i) the correctness of our rationales and (ii) the correctness of the number of reasoning steps (sentences) within the rationales. The annotators deem our rationales and their number of reasoning steps correct in 82% and 71% of samples, respectively (245/214 out of 300), reaffirming the high quality of our rationales.

4.3 Benefit of II-MMR in zero-shot stage

Overall Answer Accuracy. Table 6 presents the benefit of II-MMR in the context of the model's overall accuracy on GQA and A-OKVQA in the zero-shot setting. First, compared to the traditional CoT (BLIP2+CoT), our II-MMR$_{\text{APCOT}}$ shows superior performance (GQA: 39.08% vs. 45.79%, A-OKVQA: 36.06% vs. 49.31%). Similarly, our II-MMR$_{\text{KTPROMPT}}$ outperforms it by 6.3%. This suggests that providing the prediction (or the knowledge triplet) as input context helps the model correct wrong reasoning paths.

Second, our II-MMR$_{\text{APCOT}}$ notably outperforms the baseline (BLIP-2) on both benchmarks (e.g., A-OKVQA: 49.31% vs. 46.05%), implying that even initially *incorrect* predicted answers are beneficial as input context. Based on the empirical analysis of this case, we find that the incorrect predictions are often closely related to the correct answer. For instance, some questions ask about the object color, and the model indeed predicts the color-related answer but is incorrect (e.g., Prediction: "purple", Ground-truth: "green"). In this case, our II-MMR$_{\text{APCOT}}$, providing the prediction as a *possible* answer (cf. Top in Figure 3), may guide the model to "rethink" the correct answer (e.g., true color), eventually fixing the wrong answer. See the appendix for its qualitative examples.

Answer Accuracy over Hops. We further measure the model accuracy across different reasoning cases (hops) to better understand its reasoning capabilities (Table 6). First, aligned with the overall accuracy, our II-MMR$_{\text{APCOT}}$ notably outperforms the conventional CoT (BLIP-2+CoT) in all reasoning cases on both VQA benchmarks (e.g., 1-hop in GQA: 43.88% vs. 35.41%, $\geq$2-hop in A-OKVQA: 48.09% vs. 34.15%). We again attribute this to the benefit of incorporating the prediction into the CoT reasoning. Second, II-MMR$_{\text{APCOT}}$ consistently improves over the baseline (BLIP-2) in all reasoning scenarios of both benchmarks, demonstrating the effectiveness of our answer reasoning path (e.g., 1-hop in A-OKVQA: 51.32% vs. 46.70%). More interestingly, on the challenging reasoning cases (e.g., 2 or more-hop reasoning), our II-MMR$_{\text{KTPROMPT}}$ and II-MMR$_{\text{APCOT}}$ achieve notable gains over the baseline (e.g., GQA/A-OKVQA: 27.45% vs. 7.66% / 48.09% vs. 46.54%). In contrast, the traditional CoT achieves less gain or performs worse (GQA/A-OKVQA: 18.09% / 34.15%). This highlights the benefit of II-MMR, especially for solving complex reasoning questions.

Applicability of II-MMR. In addition to BLIP-2, we evaluate the effectiveness of our II-MMR on a more recent VLM, LLaVA-1.5 (Liu et al., 2023). As shown in Table 6, II-MMR consistently

Model	GQA				A-OKVQA		
	0-hop	1-hop	2-hop	All	1-hop	≥2-hop	All
BLIP-2	49.62	42.41	7.66	44.63	46.70	46.54	46.05
BLIP-2+CoT	44.27	35.41	18.09	39.08	37.34	34.15	36.06
BLIP-2+II-MMR$_{\text{APCOT}}$	49.69	43.88	19.36	45.79	**51.32**	**48.09**	**49.31**
BLIP-2+II-MMR$_{\text{KTPROMPT}}$	49.62	42.59	27.45	45.45	-	-	-
LLaVA-1.5+CoT	67.41	56.16	43.40	61.16	-	-	-
LLaVA-1.5+II-MMR$_{\text{APCOT}}$	**70.47**	**57.51**	**46.60**	**63.42**	-	-	-

Table 6: **Effectiveness of II-MMR over different reasoning cases in zero-shot VQA.** Our II-MMR outperforms the traditional CoT on every reasoning case. Moreover, II-MMR notably improves over the baseline (BLIP-2) in all reasoning cases on both benchmarks, suggesting the benefit of our answer reasoning path in finding correct answers. Moreover, II-MMR shows its applicability to a recent VLM, LLaVA-1.5 (Liu et al., 2023). Following Liu et al. (2023), we do not evaluate its zero-shot performance on A-OKVQA, as A-OKVQA was used during model training.

Model	A-OKVQA		
	1-hop	≥2-hop	All
BLIP-2	57.63	54.92	56.88
BLIP-2+CoT	54.65	54.40	54.58
BLIP-2+II-MMR$_{\text{APCOT}}$	**58.16**	**55.24**	**57.35**

Table 7: **Effectiveness of II-MMR$_{\text{APCOT}}$ on A-OKVQA in the fine-tuning setting.** Aligned with the zero-shot results (Table 6), our II-MMR notably outperforms two baselines, BLIP-2 and BLIP2+CoT, on every reasoning case.

Hop Increase Percentage (%)				
GQA	0-hop	1-hop	2-hop	All
	93.84	77.47	70.58	86.34

Table 8: **Hop Increase Percentage by our augmentation.** We provide the LLM with the knowledge from large-scale text corpus (Wikipedia) and make existing questions more complex (e.g., 77.47% of original 1-hop questions now have at least one more reasoning steps).

Question Type	GQA			
	0-hop	1-hop	2-hop	All
Original Q	49.62	42.41	7.66	44.63
Augmented Q	37.86	35.58	6.38	35.60

Table 9: **Zero-shot accuracy on expanded GQA questions.** The performance on augmented questions (Augmented Q) notably declines against that on original questions (Original Q), suggesting increased reasoning in the expanded questions.

improves over the traditional CoT in all reasoning cases, demonstrating its applicability to various VLMs with different architectural designs.

4.4 Benefit of II-MMR in fine-tuning stage

Besides assessing the zero-shot outcome, we evaluate II-MMR in the fine-tuning scenario. We first leverage the pre-trained BLIP-2 model with our II-MMR$_{\text{APCOT}}$ to generate a rationale. We then use this generated rationale (together with the question and the image) to fine-tune the model on A-OKVQA (See the appendix for more details). As shown in Table 7, our II-MMR$_{\text{APCOT}}$ consistently improves the baseline (BLIP-2) over different reasoning cases, including multi-hop reasoning (e.g., ≥2-hop: 55.24% vs. 54.92%). This suggests the benefit of our II-MMR$_{\text{APCOT}}$ even for fine-tuned models. Conversely, when the standard CoT is applied to BLIP-2 (BLIP-2+CoT), its performance notably degrades across all reasoning cases (e.g., ≥2-hop: 54.40%), consistent with the findings from zero-shot experiments (§4.3).

4.5 Expanding questions with more reasoning

As depicted in Table 2, the number of complex reasoning questions (i.e., 2-hop reasoning) is marginal in the GQA test-dev set, comprising only 3.73%. We thus conduct an ablation study: increasing the number of hops for each original question using a large-scale knowledge-base (e.g., Wikipedia) and evaluating the model performance on these newly expanded questions. We first extract keywords from the original question and use them as queries to retrieve their relevant information from Wikipedia. We then input this extra information (with the original question) into the LLM to increase the reasoning complexity of the question.

Figure 5 provides a detailed example of an in-context prompt for augmenting questions. Our primary objective is to increase the reasoning complexity of the original question while retaining its original answer. This enables us to conduct a more precise evaluation of how the model performance changes as the question becomes more intricate. We adopt a 5-shot in-context prompting where each example consists of seven components; "Task", "Original Question", "Original Short Answer", "Captions", "Bridge Entity", "Complex Question", and "Short Answer".

"Task: Use all captions to make the original question more complex.
Do not change the original short answer.
Do not mention the bridge entity in the complex question. If necessary, use its pronoun such as 'one' or 'object'.
Original Question: What is the policeman to the left of the truck riding?
Original Short Answer: motorcycle.
Captions: Police officers are generally charged with the apprehension of suspects and the prevention, detection, and reporting of crime, and the maintenance of public order. Some officers are trained in special duties, such as counter-terrorism or surveillance.
Bridge Entity: policeman.
Complex Question: What is the one left of the truck who is sometimes trained on counter-terrorism riding?
Short Answer: motorcycle.

Task: ...

Task: Use all captions to make the original question more complex.
Do not change the original short answer.
Do not mention the bridge entity in the complex question.
If necessary, use its pronoun such as 'one' or 'object'.
Original Question: What is hanging above the chalkboard?
Original Short Answer: Picture.
Captions: A blackboard or a chalkboard is a reusable writing surface on which text or drawings are made with sticks of calcium sulphate or calcium carbonate, known, when used for this purpose, as chalk. Blackboards were originally made of smooth, thin sheets of black or dark grey slate stone.
Bridge Entity: chalkboard.
Complex Question:"

Figure 5: **In-context language prompting to make the original question more complex.** "Bridge Entity" is a keyword extracted from the original question. "Captions" is the text snippet containing information about the bridge entity retrieved from Wikipedia. Using Wikipedia captions, we ask the LLM to increase the reasoning complexity in the original question while maintaining its original answer. We provide five in-context examples to the LLM.

Figure 6 shows a qualitative example of augmenting the original question with more reasoning. The question originally requiring 1-hop reasoning (i.e., ("surfer", "wearing", "wetsuit")) now asks for 2-hop reasoning (i.e., ("one", "wearing", "garment"), ("garment", "usedFor", "thermal protection")).

Table 8 shows that the number of hops in most original GQA questions has increased. For instance, 70.58% of original 2-hop questions now require at least one more reasoning step than previously. In total, 86.34% of original questions become more complex. We evaluate BLIP-2 on these new questions in the zero-shot setting (Table 9).

Original Q: "Is the surfer that looks wet wearing a wetsuit?" A: "Yes"
1hop reasoning: (surfer, wearing, wetsuit)

Augmented Q: "Is the one that looks wet wearing a garment for thermal protection while wet?" A: "Yes"
2hop reasoning: (one, wearing, garment), (garment, usedFor, thermal protection)

Figure 6: **Qualitative results of augmenting question.** The original question now becomes more complex with one more reasoning step.

Compared to its performance on the original questions, we observe a notable drop in performance across all reasoning cases (e.g., 7.66% vs. 6.38% in the original 2-hop), indicating increased reasoning complexity in the newly expanded questions.

4.6 Qualitative Results

Figure 7 summarizes qualitative examples provided by our II-MMR$_{\text{APCoT}}$. The generated rationale is highly relevant to the correct answer, leading the model to make the correct prediction. Moreover, our rationales entail knowledge beyond images, such as commonsense (e.g., "The clown fish is a popular kite design"), which is advantageous for A-OKVQA questions requiring external knowledge.

5 Related Work

Multi-hop Reasoning. Chain-of-thought (CoT) reasoning can be regarded as a form of multi-hop reasoning, as it involves constructing a sequence of reasoning to derive the answer. However, as indicated in Table 1, the traditional CoT method often results in erroneous information (e.g., visual hallucinations), leading to incorrect answers. Several recent works (Dhuliawala et al., 2023; Jiang et al., 2023; Chen et al., 2024; Zelikman et al., 2022) have thus attempted to enhance the CoT reasoning capabilities. CoVe (Dhuliawala et al., 2023) constructs a set of verification questions and uses them to verify the correctness of rationales. FLARE (Jiang et al., 2023) iteratively identifies flawed rationales and corrects them by utilizing relevant documents from the knowledge base. CURE (Chen et al., 2024) uses feedback from LLMs to tackle the hallucination during the generation process. Prophet (Yu et al., 2023) is related to our work as it first obtains answer candidates and then includes them in the LLM prompt for the final prediction. However, Prophet relies on separate models for answer candidate selection (VQA) and final prediction (GPT-3 (Brown et al., 2020)), which is less flexible than our single V&L model approach. Additionally, Prophet makes direct final predictions without generating rationales, thus lacking multi-hop reasoning capabilities. STaR (Zelikman et al., 2022) improves the model's reasoning capability by supervising rationale generation based on the inference results of the train sets. In contrast, our II-MMR proposes a self-correction mechanism, which uses the initial prediction as the input context and autonomously corrects it if incorrect through the reasoning steps.

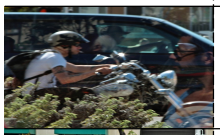
Image	Question	GT Answer	Final Prediction	Generated Rationale
	"What is in the motorcyclist's mouth?"	cigarettes	cigarettes	"The motorcyclist is smoking a cigarette while riding his motorcycle"
	"What is this place?"	market	market	"A man is selling carrots at a market"
	"What kind of fish inspired the kite design?"	clownfish	clownfish	"A clown fish is a fish with bright colors and a clown-like face. The clown fish is a popular kite design"
	"What is this game played with?"	baseball	baseball	"Baseball is a game played with a ball and a bat. The man in the picture is holding a bat."
	"What country do these planes belong to?"	usa	usa	"The planes are in a museum. The museum is in the usa"

Figure 7: **Qualtiative Results of II-MMR.** Our rationales are highly relevant to the ground-truth answer. See details in §4.6.

Explaining Reasoning on VQA. A few prior works (Li et al., 2018; Wu et al., 2020; Wu and Mooney, 2019; Vaideeswaran et al., 2022) have explained the model's reasoning capabilities in the context of VQA tasks. VQA-E (Li et al., 2018) proposes a new VQA dataset derived from VQAv2 benchmark (Goyal et al., 2017) by synthesizing explanations for original VQAv2 samples. Some prior studies (Wu et al., 2020; Wu and Mooney, 2019) leverage human textual explanations to gain further insights into the model's reasoning abilities. More recently, (Vaideeswaran et al., 2022) aims to interpret the actions of VQA models by incorporating an end-to-end explanation generation module. Conversely, we utilize the LLM with novel language promptings grounded in the answer prediction and the knowledge triplet to automatically analyze various reasoning scenarios in VQA.

Scene Graph and Knowledge Graph. A scene graph (SG) from images and a knowledge graph (KG) related to questions (Xie et al., 2022; Singh et al., 2023) are alternative ways to find the reasoning path to the answer. However, compared to rationales generated from LLM-based V&L models, SG and KG usually provide restricted visual semantic details in explaining the reasoning. This limitation arises as their graph generators (Zheng et al., 2023; Schuster et al., 2015) often fail to capture diverse visual entities or semantic relations.

6 Conclusion

We propose II-MMR to identify and improve multi-hop reasoning for VQA. II-MMR introduces two novel language promptings, an answer prediction-guided CoT prompt and a knowledge triplet-guided prompt, to generate a high-quality reasoning path to reach the answer. II-MMR utilizes this path to identify different reasoning scenarios in VQA benchmarks and consistently improves across all reasoning cases with a particular emphasis on complex reasoning questions.

Limitations

In this work, we propose II-MMR, a novel method to improve the reasoning capabilities of V&L models for VQA. We conduct a small-scale human study (involving 3 annotators) to assess the quality of our rationales. The evaluation may be subjective among annotators due to the nature of the language. For instance, each annotator may have different opinions about the number of reasoning steps required for the same question. We plan to expand the scale of the human study to mitigate this issue.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of ICCV*.

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. 2023. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *NeurIPS*, pages 1877–1901.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2023. Pali: A jointly-scaled multilingual language-image model. In *ICLR*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Measuring and improving chain-of-thought reasoning in vision-language models. In *NAACL*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *JMLR*, 25(70):1–53.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*.

Drew A Hudson and Christopher D Manning. 2019. 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*, pages 6700–6709.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *EMNLP*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

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

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80.

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

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *EMNLP*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of EMNLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rakesh Vaideeswaran, Feng Gao, Abhinav Mathur, and Govind Thattai. 2022. Towards reasoning-aware explainable vqa. In *NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML)*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research (TMLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jialin Wu, Liyan Chen, and Raymond J Mooney. 2020. Improving vqa and its explanations\\by comparing competing explanations. *arXiv preprint arXiv:2006.15631*.

Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. 2022. Visual clues: Bridging vision and language foundations for image paragraph captioning. *Advances in Neural Information Processing Systems*, 35:17287–17300.

Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. 2023. Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. In *CVPR*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *NeurIPS*.

Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. 2023. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792.

Appendices

In this appendix, we provide details omitted in the main text.

- Appendix A: More details about dataset and training (cf. §3).
- Appendix B: Qualitative results of incorrect prediction (cf. §4.3).
- Appendix C: Fine-tuning details (cf. §4.4).
- Appendix D: Additional Results (cf. §3).

A More dataset and training details

Dataset. We provide further details about the datasets used in our experiments. GQA (Hudson and Manning, 2019) is one of the popular VQA benchmarks comprising various visual compositional reasoning questions. GQA first obtains relations and attributes of visual objects from the scene graph (Krishna et al., 2017) and utilizes them to generate VQA questions based on a pre-defined question engine. Followed by BLIP-2 (Li et al., 2023), we use the official GQA train/test-dev splits for our experiments. A-OKVQA (Schwenk et al., 2022) contains diverse VQA questions requiring real-world knowledge beyond the image, including commonsense and knowledge bases. Concretely, A-OKVQA has 25K questions, each offering both multiple-choice and direct-answer options. For our studies, we select the direct-answer option. We utilize the A-OKVQA train/validation splits, which consist of 17.1K/1.1K samples, respectively.

Image	Question	GT Answer	Initial Prediction	Generated Rationale	Final Prediction
	what is the same color as the most abundant color here?	Green	Purple	The most abundant color here is green. Therefore, short answer: green	Green

Figure 8: **Qualtiative Results of our II-MMR with incorrect prediction.**

Training. We mainly use BLIP-2 as the V&L model for our experiments. Concretely, we opt for the largest BLIP-2 model, which features the ViT-g/14 vision encoder from EVA-CLIP (Fang et al., 2023), coupled with FlanT5-XXL (Chung et al., 2024), an encoder-decoder-based LLM. BLIP-2 incorporates a transformer-based bridge module that connects the vision encoder to the LLM. We focus on training the LLM and the bridge component while keeping ViT frozen. Our training configuration entails a batch size of 16, a learning rate of 1e-5, a beam size of 5, a maximum sequence length of 512, and an image resolution of 490 for ten epochs. We utilize eight A100 48GB GPUs for both training and inference.

B Qualitative results of incorrect prediction

Figure 8 shows an example that our II-MMR$_{\text{APCoT}}$ with the initial incorrect prediction leads the V&L model to make a correction, aligning with its quantitatively superior performance compared to the baseline (BLIP-2) (e.g., A-OKVQA All: 49.31% vs. 46.05% in Table 6).

C Fine-tuning details

We provide details about fine-tuning our II-MMR$_{\text{APCoT}}$ on A-OKVQA. As mentioned in §4.4, we initially utilize the same pre-trained BLIP-2 model used for zero-shot tasks, along with our II-MMR$_{\text{APCoT}}$, to generate an answer-related rationale. The main motivation for using the pre-trained model for rationale generation (instead of selecting a fine-tuned one on A-OKVQA) is that once the model undergoes fine-tuning, it loses its capability to generate rationales and shifts its primary focus to predicting answers directly. We thus deliberately select the pre-trained model for the effective rationale generation process. After obtaining the rationale, we prepend it (with the question and the image) to the answer-trigger prompt (e.g., "Therefore, short answer:") and fine-tune BLIP-2 on A-OKVQA using this prompt to predict the answer.

D Additional results

Accuracy of II-MMR$_{\text{APCoT}}$ with the ground-truth answer as input context. To accurately measure the number and the types of reasoning in A-OKVQA (Schwenk et al., 2022), we utilize the ground-truth answer as the input context for II-MMR$_{\text{APCoT}}$, instead of initial answer prediction, which is the default setting. We observe a significant improvement in A-OKVQA performance when utilizing our II-MMR with the ground-truth answer compared to traditional CoT (69.34% vs. 36.06%), again supporting the high quality of our reasoning path for analyzing different reasoning scenarios (cf. §4.1).