ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models

Kaiwen Zhou[1], Kwonjoon Lee[2], Teruhisa Misu[2], and Xin Eric Wang[1]

[1]University of California, Santa Cruz
[2]Honda Research Institute

Abstract

In our work, we explore the synergistic capabilities of pre-trained vision-and-language models (VLMs) and large language models (LLMs) on visual commonsense reasoning (VCR) problems. We find that VLMs and LLMs-based decision pipelines are good at different kinds of VCR problems. Pre-trained VLMs exhibit strong performance for problems involving understanding the literal visual content, which we noted as visual commonsense understanding (VCU). For problems where the goal is to infer conclusions beyond image content, which we noted as visual commonsense inference (VCI), VLMs face difficulties, while LLMs, given sufficient visual evidence, can use commonsense to infer the answer well. We empirically validate this by letting LLMs classify VCR problems into these two categories and show the significant difference between VLM and LLM with image caption decision pipelines on two subproblems. Moreover, we identify a challenge with VLMs' *passive* perception, which may miss crucial context information, leading to incorrect reasoning by LLMs. Based on these, we suggest a collaborative approach, named **ViCor**, where pre-trained LLMs serve as problem classifiers to analyze the problem category, then either use VLMs to answer the question directly or *actively* instruct VLMs to concentrate on and gather relevant visual elements to support potential commonsense inferences. We evaluate our framework on two VCR benchmark datasets and outperform all other methods without in-domain fine-tuning.

1 Introduction

The problem of visual commonsense reasoning (VCR) (Zellers et al., 2019; Hessel et al., 2022; Schwenk et al., 2022) expands upon the traditional visual question answering (Antol et al., 2015; Goyal et al., 2017). VCR requires machines to understand complex visual scenes, extract crucial visual content, and utilize commonsense knowledge for drawing novel conclusions that go beyond the explicit information present in the image. Previous methods have utilized pre-trained large language models and pre-trained or fine-tuned vision-language models (Hu et al., 2022; Shao et al., 2023; You et al., 2023) to solve VCR problems in few-shot or fine-tuned setting.

However, some open questions exist on how VLMs and LLMs can efficiently and effectively collaborate to solve these VCR problems. Firstly, what roles do LLMs and VLMs play in solving VCR problems with their different capabilities? Secondly, how do we maximize their capabilities to solve the VCR problems without in-domain fine-tuning?

To answer these two questions, as shown in Figure 1, we first find that VLMs themselves can solve the problems requiring the model to recognize various low-level visual patterns and understand high-level concepts like actions, events, and relations indicated by those visual patterns. In the meanwhile, solving problems that require the model to deduce conclusions or form explanations based on visual observation relies more on the commonsense reasoning capabilities of LLMs. This kind of problem requires a broad array of commonsense knowledge about the world, including cause-and-effect relationships, intentions, and mental states (Sap et al., 2020). To validate this finding, we first note these two kinds of VCR problems as *visual commonsense understanding (VCU)* and *visual commonsense inference (VCI)*. Then, we instruct LLMs to classify VCR question into these two categories. We empirically find that, for VCU problems, VLMs like BLIP2 can achieve better results than LLM+caption pipeline (Yang et al., 2022) with their visual understanding capabilities, and for VCI problems, the LLM+caption pipeline is better.

In the meanwhile, we observe that image captions provided by VLMs often lack crucial contextual information necessary for answering questions.

Findings of the Association for Computational Linguistics: ACL 2024, pages 10783–10795
August 11-16, 2024 ©2024 Association for Computational Linguistics

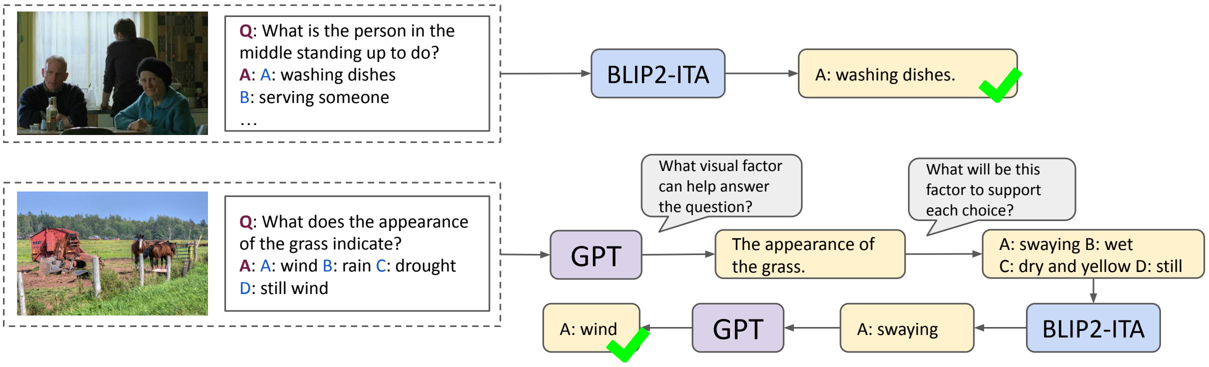

Q: What is the person in the middle standing up to do?
A: A: washing dishes
B: serving someone
…
BLIP2-ITA
A: washing dishes.
Q: What does the appearance of the grass indicate?
A: A: wind B: rain C: drought
D: still wind
GPT
What visual factor can help answer the question?
The appearance of the grass.
What will be this factor to support each choice?
A: swaying B: wet
C: dry and yellow D: still
BLIP2-ITA
A: swaying
GPT
A: wind

Figure 1: Two examples demonstrating different kinds of visual commonsense reasonings require different model capabilities. **Upper: Visual commonsense understanding (VCU)** requires the model to understand high-level concepts and attributes such as actions, events, relations, *etc*, which pre-trained VLMs can achieve via image-text alignment (ITA). **Lower: Visual commonsense inference (VCI)** requires the model to generate conclusions or explanations based on input image. Overlooking visual clues can result in erroneous conclusions. LLMs steer VLMs in discovering vital visual cues for answer support. The LLM employs the top ITA-scored visual clue (e.g.,“It is cloudy.") to perform commonsense inference.

This poses a particular challenge for commonsense inference problems, as inferences are often defeasible given additional context (Choi, 2022). To illustrate this issue, consider the example depicted in Figure 1 (bottom). At first glance, it may appear that there’s nothing noteworthy beyond horses on a grassy farm, leading one to select “D: still wind" as an answer. However, upon closer examination of the swaying grass, we must revise our conclusion to “A: wind." Existing perception modules, including VLMs, operate in a feed-forward manner and cannot adjust their perception based on a high-level understanding or inference. To address this, we propose instructing LLMs to intervene with VLMs in cases where they are uncertain about inference, indicating a lack of sufficient visual evidence. This intervention would guide VLMs to focus on specific *visual factors*, such as weather or emotions, to support commonsense inferences.

Based on these findings, we propose the **ViCor** framework, which employs the following components: (1) LLMs functioning as problem type classifiers (VCU and VCI), VLM commanders for directing VLMs based on reasoning, and visual commonsense reasoners. (2) Pre-trained VLMs are responsible for visual recognition and understanding. Communication between LLMs and VLMs occurs through text, such as image captions. On VCR (Zellers et al., 2019) and A-OKVQA (Schwenk et al., 2022), our method achieves state-of-the-art results among methods *without* supervised in-domain fine-tuning.

2 Related Work

Visual Commonsense Reasoning Visual Commonsense Reasoning (VCR) (Zellers et al., 2019; Hessel et al., 2022; Schwenk et al., 2022) is an emerging research area that aims to endow AI models with a human-like understanding and reasoning of visual scenes. The goal is to understand high-level concepts such as events, relations, and actions and infer unobservable aspects such as intents, causal relationships, and future actions. The VCR task was introduced by Zellers et al. (2019). Further, more datasets focused on more types of reasoning were proposed (Park et al., 2020; Hessel et al., 2022; Schwenk et al., 2022). Most methods treat VCR as an image-text alignment problem, where they encode the commonsense inference and the visual input, then predict the alignment score of the image-text pair (Zellers et al., 2019; Chen et al., 2020; Zellers et al., 2022; Hessel et al., 2022). Although achieving impressive performance, the generalizability of these methods is limited by supervised training.

Recently, several works have leveraged large language models for visual commonsense reasoning (Hu et al., 2022; Shao et al., 2023; You et al., 2023). However, (Hu et al., 2022; Shao et al., 2023) require some VLMs trained on the datasets to provide visual information. (You et al., 2023) use LLMs to decompose the main problem and use VQA models to acquire visual information. Following these works, we take one step further and systematically study the strength of pre-trained VLMs and the reasoning abilities of LLMs on visual commonsense reasoning problems. We then propose a framework to efficiently and effectively leverage the advantages of both models.

Large Language Models for Vision-and-Language Tasks Benefiting from the rich knowledge in LLMs, they have been used for various vision-and-language tasks in a zero-shot or few-shot manner. Yang et al. (2022); Hu et al. (2022); Shao et al. (2023) leverage LLMs for OK-VQA task (Marino et al., 2019) by feeding the caption, question, candidate answers by VQA models, etc. to GPT3, and prompt the GPT3 to answer the question with its pre-trained knowledge. More recently, with the discovery of LLMs' tool using ability (Yao et al., 2023; Schick et al., 2023), LLMs were equipped with various visual tools (Gupta and Kembhavi, 2023; Dídac et al., 2023; Shen et al., 2023; Lu et al., 2023; Wu et al., 2023) and achieved significant performance in Compositional Visual Question Answering, Science Question Answering tasks (Suhr et al., 2018; Hudson and Manning, 2019; Lu et al., 2022). Concurrently, (Wu and Xie, 2023) proposes to use LLM reasoning to help concentrate on the right part of the image to answer visual questions, which share a similar spirit with our proposed method. Our work studies a complex and challenging problem – visual commonsense reasoning, which requires different capabilities of LLMs. In our method, we fully leverage LLM capabilities to conduct reasoning for problem classification, visual information query, and commonsense reasoning.

3 Visual Commonsense Reasoning

3.1 Problem Categorization

We first illustrate our categorization of VCR problems to distinguish the capabilities of VLMs and LLMs.

Visual Commonsense Understanding The visual commonsense understanding (VCU) problem requires the model to judge if a text T describing a concept or an attribute aligns with the image I:

$$e = F(I, T) \tag{1}$$

where e stands for evaluation of T by model F. To answer these questions, the model needs to be able to map the low-level visual observations, such as objects and spatial relations to various high-level visual concepts and attributes, such as landmarks, actions, events, and relations.

Visual Commonsense Inference The visual commonsense inference (VCI) problem usually requires the model to evaluate the plausibility of an inference about the image. Besides understanding the literal content in the image as in VCU, evaluating the inferences T in VCI problems needs involves drawing novel conclusions or explanations from these visuals, often using (non-visual) commonsense knowledge, based on some visual observations $\{o_i\}$ derived from the image:

$$e = F(\{o_i\}, T) \tag{2}$$

Here, o_i could be some visual observations or high-level visual commonsense understanding. Examples of non-visual commonsense knowledge could be the purpose of an object, people's opinions about an object, potential future events, etc. We show the performance difference between VLMs and LLMs-based decision models on two sub-problems in Table 1, which will be illustrated in Sec. 6.1.

3.2 Problem Formulation

Both categories of visual commonsense reasoning tasks share a common formulation. In visual commonsense reasoning, the input consists of two parts: an image denoted as I and a multiple-choice question input represented as q, c_i, where q corresponds to the question, and c_i stands for the i-th answer choice. The model needs to choose the choice c_i that is most likely to be true based on the image I.

4 The ViCor Framework

To enable more effective collaboration between LLMs and VLMs, as shown in Figure 2, we design an approach that involves a multi-step process. First, a pre-trained LLM takes the initial perception result (*i.e.*, image caption), a question-answer pair, and instructions as input to evaluate potential answer candidates. Then, if the LLM is not confident about its reasoning, it will select to use VLMs to directly answer the question or to guide the VLMs to collect target visual information for re-evaluation.

4.1 Large Language Models as VCR Reasoner

Evaluating answer choices in VCR requires drawing new conclusions based on commonsense knowledge, which LLMs excels at (Anil et al., 2023). On the other hand, pre-trained vision-and-language

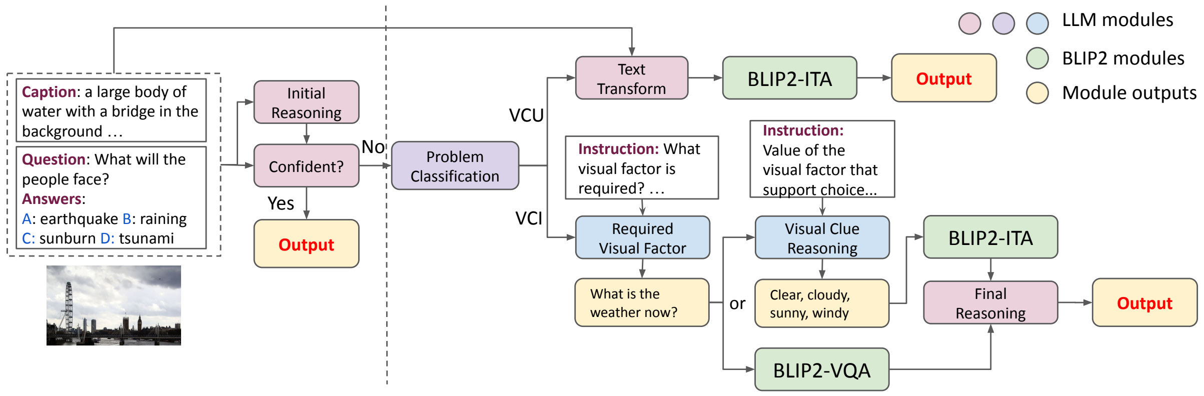


Figure 2: **Our ViCor framework.** Given a visual commonsense reasoning problem and a caption, our framework will leverage LLM to perform initial reasoning and confidence check. If the reasoning is not confident, the LLM will perform problem classification and acquire visual information according to the problem type. $*$Note that the final reasoning takes the question and the caption as input as well.

models have exhibited a capability for visual understanding, such as image captioning and image-text alignment, with a demonstrated ability to generalize across various datasets (Li et al., 2023). Therefore, we decided to harness the strengths of vision-and-language models for visual understanding and the capabilities of large language models for evaluating answer candidates in the context of visual commonsense reasoning.

Captioning serves as a fundamental unsupervised pre-training task and the most generalized capabilities of pre-trained VLMs, which capture the most salient information from an image. Therefore, we first prompt the LLMs to take the caption of the image C_I as the initial information and perform chain-of-thought reasoning on the question:

$$r_1 = LLM(\{c_i\}, q, C_I). \tag{3}$$

The reasoning result r_1 includes both intermediate reasoning steps and the final answer. However, it's important to note that the image caption may not encompass all the relevant information within the image, potentially omitting critical contextual details essential for answering the question. In such cases, it becomes necessary to gather additional relevant visual observations from the image. Before this, we must first judge whether there is a lack of supportive visual evidence that would allow us to make a confident decision. As in Figure 2, we let the LLM take the initial reasoning r_1 and the history prompt as input to judge if current visual information adequately supports the decision. If it does, the model will directly output the result. Conversely, if there is a lack of sufficient evidence, the model will progress to the second stage, where it will actively seek additional visual evidence.

4.2 Large Language Models as VCR Problem Classifier

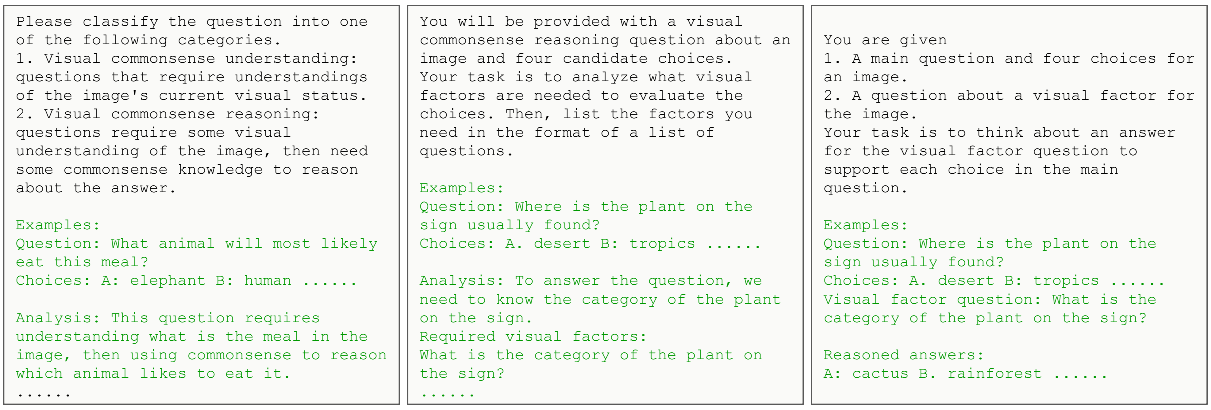
As mentioned before, we found the capabilities of VLMs are suitable for solving VCU problems, and LLMs are more capable of solving VCI problems. Therefore, we propose to leverage VLMs in distinct manners when facing different problem types. However, we first need to know what given a visual question, To this end, we first prompt the LLM to classify the problem into two categories. To achieve this, we provide the definitions of these two categories in the prompt. Additionally, we include a set of manually annotated in-context examples to aid in problem classification, where the questions of in-context examples are selected from the training set. Figure 3 illustrates the prompt.

4.3 Large Language Models as VLM Commander

The pre-training dataset of vision-and-language models contains millions to billions of image-text pairs. Therefore, VLMs have learned the mapping between visual features and the high-level commonsense concept well. In light of this, for *visual commonsense understanding (VCU)* problems, we leverage pre-trained VLM in a zero-shot manner. Specifically, for each choice c_i, we first instruct the LLM to transfer it and the question to a declarative sentence with instruction and in-context examples:

$$s_i = LLM(q, c_i) \tag{4}$$

For instance, for the question `What will the people face?` and the choice `earthquake`, we



```
Please classify the question into one
of the following categories.
1. Visual commonsense understanding:
questions that require understandings
of the image's current visual status.
2. Visual commonsense reasoning:
questions require some visual
understanding of the image, then need
some commonsense knowledge to reason
about the answer.

Examples:
Question: What animal will most likely
eat this meal?
Choices: A: elephant B: human ......

Analysis: This question requires
understanding what is the meal in the
image, then using commonsense to reason
which animal likes to eat it.
......
```

```
You will be provided with a visual
commonsense reasoning question about an
image and four candidate choices.
Your task is to analyze what visual
factors are needed to evaluate the
choices. Then, list the factors you
need in the format of a list of
questions.

Examples:
Question: Where is the plant on the
sign usually found?
Choices: A. desert B: tropics ......

Analysis: To answer the question, we
need to know the category of the plant
on the sign.
Required visual factors:
What is the category of the plant on
the sign?
......
```

```
You are given
1. A main question and four choices for
an image.
2. A question about a visual factor for
the image.
Your task is to think about an answer
for the visual factor question to
support each choice in the main
question.

Examples:
Question: Where is the plant on the
sign usually found?
Choices: A. desert B: tropics ......
Visual factor question: What is the
category of the plant on the sign?

Reasoned answers:
A: cactus B. rainforest ......
```

Figure 3: Three simplified prompt examples demonstrating how we define prompts to classify the problem (**left**), reason visual factors (**middle**), and think about visual observations regarding visual factors (**right**).

will transform them to `The people will face earthquake`. Then, we feed s_i and the image I to the pre-trained VLM to calculate the image-text alignment score. Following (Li et al., 2023), we use the sum of ITM and ITC scores to compare choices:

$$S_i = ITM(I, s_i) + ITC(I, s_i) \tag{5}$$

We will directly take the choice with the highest score as the final output.

For the *visual commonsense inference (VCI)* problems, the model needs to acquire related visual observations and use relevant commonsense knowledge to reason about the answer. Some crucial visual observations are often neglected in the descriptions of the image. Therefore, as in Figures 2 and 3, we first prompt the LLMs to think about some *visual factors* f_j that influence the answer to the question, like 'the action of the person', 'the interaction between people', etc. Then, we could acquire the visual observation of the visual factor in the image with a visual question-answering model by asking a question about the visual factor:

$$o_j = VQA(I, f_j) \tag{6}$$

where o_j is the answer to the question which we call *visual clue*. However, the answer of VQA does not consider the context of the main question and therefore may lack the most related information. To better leverage the contextualized reasoning capabilities of LLMs, we further propose to prompt the LLM to reason the potential instantiations of the visual factors that can support the choices as in Figure 3:

$$o_{ij} = LLM(f_j, c_i, q) \tag{7}$$

For instance, when f_j is "category of the plant," the potential values for o_{ij} may include specific plant names like "cactus." Then, we could leverage the image-text matching (ITM) and image-text contrastive (ITC) functions of pre-trained VLMs to select the observation that most align with the image among the observations for each choice i:

$$o_j = o_{kj} \tag{8}$$

$$\text{where } k = argmax_i\{ITM(o_{ij}, I) + ITC(o_{ij}, I)\} \tag{9}$$

Finally, we append the *visual clues* $\{o_j\}$ after the caption as extra information for LLM to perform final reasoning:

$$r_2 = LLM(\{c_i\}, q, C_I, \{o_j\}) \tag{10}$$

5 Experiments

5.1 Datasets

We mainly evaluate our approach on two datasets focused on visual commonsense reasoning: VCR (Zellers et al., 2019) and AOKVQA (Schwenk et al., 2022).[1] Both datasets formulate visual commonsense reasoning as 4-choice QA problems about an image, containing various visual commonsense understanding and inference problems. VCR dataset focuses on human-centric visual commonsense reasoning problems. In contrast, A-OKVQA dataset requires various commonsense knowledge about common objects and events in daily life. For A-OKVQA, we use the validation set with 1145 examples. For VCR

[1] We provide the result on two direct-answer VCR datasets – OKVQA and the direct-answer version of AOKVQA in Appendix. A.2.

Table 1: Ablations on the effect of LLMs and VLMs on VCR (Zellers et al., 2019) and A-OKVQA (Schwenk et al., 2022) datasets. We use GPT-3.5-turbo-0613 for LLM-based methods. *Orig means using the declarative sentences transformed by LLM (Eq.5). *Clue means using the clues generated by LLM for image-text alignment (Eq.11). All numbers indicate accuracy (%). "Conf" indicates the samples where the LLM-Caption baseline shows confidence in its initial reasoning, while "!Conf" indicates cases where it lacks confidence.

Decision Model	Visual Info	AOKVQA				VCR			
		VCU		VCI		VCU		VCI	
		Conf	!Conf	Conf	!Conf	Conf	!Conf	Conf	!Conf
BLIP2-Pretrain	Orig*	76.5	66.3	56.5	50.9	70.0	56.3	59.2	47.4
	LLM Clue*	74.4	63.0	60.2	56.1	70.6	56.7	63.3	49.2
LLM	Caption	78.9	55.1	85.2	50.9	75.3	46.6	65.3	41.9
	Caption + VQA Clue	77.5	56.2	82.4	54.9	75.9	51.9	65.3	47.3
	Caption + LLM Clue	79.2	65.6	81.5	64.2	72.9	58.1	57.1	52.9
Num. of Examples		289	575	108	173	170	1779	49	1002

dataset, we randomly sample 3000 / 26534 examples from the validation set for the ablation study, and sample 500 examples to compare with other methods due to the GPT4 API cost. We divide the image from left to right into three bins and name the person depending on which bin they are located in when feeding text to VLMs and LLMs, similar to (You et al., 2023). The performance of both datasets is evaluated by accuracy.

5.2 Implementation Details and Baselines

In our experiments, we use GPT-3.5-turbo-0613 and GPT-4-0613 as the LLMs for reasoning. To ensure reproducibility, we set the temperature of the LLMs to 0. For image captioning, we employ LLAVA-7B-v1.1. Furthermore, we use the pre-trained BLIP2 model for image-text alignment and BLIP2-FlanT5 XL for visual question answering. The number of in-context examples used in the prompts shown in Figure 3 is 6, 1, and 3, respectively. All the questions in the in-context examples are from the training set.

We implement the following training-free baselines for comparison: **(1) BLIP2-Pretrain** (Li et al., 2023): We use the pre-trained BLIP-2 model directly to perform image-text alignment on both datasets. On both datasets, we utilize GPT-3.5-turbo-0613 to transform the questions and choices into declarative sentences and feed them to the BLIP-2 model to calculate the image-text alignment score. We select the choice with the highest alignment score as the answer. **(2) IdealGPT** (You et al., 2023): It prompts LLMs to break down the question and iteratively query a VQA model to answer sub-questions for visual reasoning. In our experiments, we employ the original source code of IdealGPT while utilizing the *same* version of LLM and VLMs for caption, VQA, and reasoning as our method.

6 Results and Analysis

6.1 Ablation Study

We conduct ablation studies about VLMs and LLMs collaboration on VCR and AOKVQA datasets. Results are shown in Table 1.

How do VLM and LLM compare on visual commonsense reasoning? By comparing the first row and the third row in Table 1, we can validate our hypothesis that VLMs perform well at VCU problems and LLMs help VCI problems better. We observe that, in VCU problems, the VLMs perform significantly better than LLM reasoning based on the caption on both datasets, with an average accuracy of 63.6% vs. 56.0%. While on VCI problems, LLM+caption performs better on average at 53.6% vs. 50.5%. We could also observe that BLIP2 has a significant performance gap between the two kinds of problems while LLM performs similarly.

How do visual factors and LLM clue reasoning help visual commonsense reasoning? We validate the effectiveness of visual factors reasoning and LLM clue reasoning on both BLIP2-Pretrain and LLM-based decision paradigms. Here, we describe how we adapt the clue generation method (as in Eq. 7) for BLIP2-Pretrain decision paradigm: we first prompt the LLM to generate the required visual factors f_j, then generate visual clues o_{ij} of these factors that can support each choice i. When applying the clues to BLIP2-Pretrain, we take the

Table 2: Comparison between ViCor and other methods on VCR Q→A task. * Results on full validation set. † CoT indicates the same setting as 'Caption' baseline in Table. 1: given caption and perform chain-of-thought reasoning.

	Method	Acc.(%)
Sup.	R2C (Zellers et al., 2019)	67.3
	*MERLOT (Zellers et al., 2021)	79.4
ICL	BLIP2-Pretrain (Li et al., 2023)	51.2
	GPT-3.5	
	†CoT	43.8
	IdealGPT (You et al., 2023)	47.9
	ViCor (ours)	55.4
	GPT-4	
	CoT	57.8
	ViCor (ours)	59.8

Table 3: Comparison between ViCor and other methods on A-OKVQA dataset. *Both PromptCap and Prophet trained VLMs on A-OKVQA dataset as part of the module. Sup. indicates supervised methods, and ICL means methods using in-context learning.

	Method	Acc.(%)
Sup.	GPV-2 (Kamath et al., 2022)	60.3
	*PromptCap (Hu et al., 2022)	73.2
	*Prophet (Shao et al., 2023)	76.4
	InstructBLIP (Dai et al., 2023)	81.0
ICL	BLIP2-Pretrain (Li et al., 2023)	65.6
	GPT-3.5	
	CoT	63.3
	ViCor (ours)	70.9
	GPT-4	
	CoT	70.3
	AssistGPT (Gao et al., 2023)	74.7
	ViCor (ours)	75.6

average of the image-text alignment scores within the same choice as the image-text alignment score for the choice i:

$$S_i = \frac{1}{n}\sum_{j}(ITM(I, o_{ij}) + ITC(I, o_{ij})) \quad (11)$$

where n is the number of required visual factors determined by LLM. The choice with the highest score will be selected.

From Table 1, we can first find that visual factors and visual clues are less helpful in **VCU** problems. On VCU problems, besides directly taking the concept being asked by the original question as the visual factor. The model will also consider low-level visual features as visual factors for the question. For example, for the question `What is the event in the image`, and the choice `dinner`, the visual factor could be `objects in the image`, and the reasoned visual clues could be `plates with food on the table`.

On BLIP2-Pretrain, using clues for image-text alignment is not better than using directly transferred declarative sentences. This validates that BLIP2 can already align visual features with different concepts well. However, introducing visual factors and observations as extra context improves performance on LLM reasoning, especially when the LLM is not confident about its initial judgment with only caption as context. In this case, the performance of LLM reasoning ('Cap + Clue' in Table 1) is comparable with pre-trained BLIP2.

For **VCI** problems, visual factors and visual clue generations help both reasoning paradigms. First, the improvement in the BLIP2-Pretrain paradigm validates that **(1)** pre-trained BLIP2 cannot well-align statements that go beyond literal visual content, requiring commonsense inference; **(2)** LLM can reason about the visual factors that may contribute to supporting candidate commonsense inferences, and guide the VLM to focus on relevant factors accordingly. Second, the improvement in the LLM reasoning paradigm shows that LLM clues successfully provide subtle details of the scene that are crucial for solving the problem. Third, visual clues reasoned by LLM are better than VQA as the visual information provider. There are mainly two reasons. First, the pre-trained VLM sometimes could not understand or correctly answer the question due to the lack of language alignment. Second, the VQA model lacks the main question as the context and may not get the intention of the visual factor. Therefore, it may produce irrelevant answers. We provide examples to further illustrate these in Section 6.3.

How to determine the reasoning process based on confidence and problem category for better LLM and VLM collaboration? When deciding the reasoning process, we need to consider both the performance and efficiency, evaluating by the number of LLM calls. From Table. 1, we can observe that when the LLM is confident about its initial reasoning, the performance is the best or almost the best on both VCU and VCI problems. Therefore, using LLM+caption is the best choice. When

Question	What is person in the middle standing up to do?	Why is woman holding umbrella?	What is the appearance of the grass indicating?
Choices	A: washing dishes B: serve someone	A: block sun B: repelling rain C: to dance	A: wind B: rain C: drought D: still wind
LLAVA Caption	A man and a woman sitting at a table in a restaurant. The man is holding a bottle.	A group of people sitting at a table outside, with a woman sitting at the end under an umbrella.	A field with a tractor, a horse, and a cow. The tractor is parked in the grass.
Initial Evaluation	Given it's in a restaurant and the man is holding a bottle, it is likely that he is serving someone.	Since the caption does not mention any bad weather such as rain, the umbrella is to block sun.	The animals may not stand in rain or drought. The grass seems not moving – no strong wind.
Category	VCU	VCI	VCI
BLIP2-ITA results	A	A	D
Visual factors	N/A	The weather	Appearance of the grass
LLM clues	N/A	A: sunny B: rainy C: sunny	A: sway B: wet C: wither D: motionless
VQA result	N/A	rainy	It is a grassy field
Final result	A: washing dishes	B: repelling rain	A: wind
Ground truth	A	B	A

Figure 4: **Qualitative examples.** All the examples are in the case of initial reasonings are not confident. **Left:** An example in the **VCR** dataset, where the ITA corrects the initial reasoning. **Middle:** An example in the **A-OKVQA** dataset, where the LLM corrects the initial reasoning after giving the observation of the visual factor. **Right:** An example in the **A-OKVQA** dataset, where the reasoned clue provides more useful information than VQA.

the LLM is not confident about its initial reasoning on VCI problems, LLM+Caption+LLM clue significantly outperforms other decision paradigms. On VCU problems, we can observe that the performance of BLIP2 is similar to LLM+Caption+LLM clue. However, the LLM+Caption+LLM clue requires five LLM calls, which is three times more than using BLIP2. Therefore, using BLIP2-ITA is the best choice in this case.

6.2 Main Results

VCR The results on VCR dataset are in Table 2. Our method achieves the best result compared with other training-free methods. Specifically, our method outperforms IdealGPT (You et al., 2023) since it is able to leverage the visual understanding abilities of VLMs more effectively by considering the types and definitions of problems. However, we notice that there is still a significant gap between ICL methods and methods with supervised training. This could be due to the loss of information in approximating the naming and labeling of the persons mentioned in Section 5.1.

A-OKVQA On A-OKVQA dataset, on both GPT models, our method can improve on chain-of-thought baseline by a significant margin. Compared with concurrent method AssistGPT (Gao et al., 2023), which utilizes GPT4 to call more visual tools such as object detection (Liu et al., 2023), text detection, and region grounding (Wang et al., 2022), our method with only BLIP2 and LLAVA can achieve better results. Meanwhile, we can observe that our method ViCor, without any training on the dataset, can achieve results close to the best supervised methods. This shows that our analysis and modeling for visual commonsense reasoning makes our framework tackle the VCR problems more efficiently.

6.3 Qualitative Examples

In Fig. 4, we demonstrate several qualitative examples. The left example shows a case where the problem is classified as VCU, and the BLIP2-Pretrain selects the correct answer. The middle example presents a case where the initial evaluation is incorrect, and both the VQA and clue reasoning methods give the correct observation for the visual factor 'weather', based on which the LLM selects the correct answer. The BLIP2-Pretrain here selects 'block sun' due to the lighting condition of the image. The example on the right demonstrates a case when the LLM reasoned answer is better than the answer generated by the VQA model. Here, the VQA does not understand the intention of the visual factor without the context of the main question. The LLM reasoned answer, however, can provide the most relevant information to the question and help the final reasoning. The BLIP2-Pretrain fails here due to the textual similarity between 'wind' and 'still wind'.

7 Conclusion

In this work, we study the collaboration of pre-trained vision-language models and large-language models on a complex problem – visual commonsense reasoning (VCR). We analyze and validate the distinct advantages of LLMs and VLMs by testing them on two different types of VCR problems. Based on this, we propose the ViCor framework that efficiently uses the visual understanding capabilities of VLMs and commonsense reasoning capabilities of LLMs to overcome the challenges in VCR. The experiment results validate the effectiveness of our framework. We believe our study can provide insights into the roles and the collaboration of LLMs and VLMs in vision and language problems.

8 Limitation and Potential Risk

In our framework, we use text as the communication medium between LLMs and VLMs. The loss of visual details caused by captions may be hindering certain scenarios, and thus, our method lags behind supervised best-performing methods. Future work could explore incorporating our designs into an end-to-end fine-tuning approach. Large language models (LLMs) are a core component of our framework. Therefore, our method may inherit the potential risks from LLMs, such as hallucination and potentially offensive language.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments and feedback.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Yejin Choi. 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155.

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

Surís Dídac, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

DiFei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *Arxiv*.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *ECCV*.

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

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.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. In *European Conference on Computer Vision*, pages 662–681. Springer.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In *ICML*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

A Additional Results

A.1 Results on More LLM Decoding Configurations

To validate the robustness of our method, we ran the experiments on the VCR dataset with more decoding configurations using LLM + Caption + LLM Clue decision branch. Specifically, we ran on two more LLM decoding temperatures 0.1 and 0.2, and used different in-context examples for the prompt in Fig.3 (right) to guide the LLM to think about observations for visual factors based on candidate choices. From the results in Table. 4, we can observe that different decoding configurations influence the results by a small margin and do not affect the main conclusions.

A.2 Results on Open-Ended VCR

We adapt our method and baselines to two open-ended VCR datasets, the OKVQA (Marino et al., 2019) dataset and the open-ended version of A-OKVQA. For OKVQA, we use the full validation set, which contains 5046 examples. The results are in Table. 5. In these experiments, We use GPT-3.5-Turbo for LLM modules. We use the Caption+VQA clue version of our method in Figure 1 to tackle unconfident VCI problems. As shown above, our framework can still leverage the advantage of both VLMs and LLMs to achieve better results thanks to the better collaboration between them, e.g., VLMs utilization based on problem classification and active visual information acquisition.

A.3 Alternative Visual Information Source

Besides the general caption, there are other alternative visual information sources for LLMs. To better demonstrate the effectiveness of our active visual information acquisition approach, we use dense captions to provide visual information instead of general captions on AOKVQA dataset. For dense captions, we use a state-of-the-art dense caption model GRiT (Wu et al., 2022). The results are shown in Table. 6.

We find that dense captions have a significant negative impact on performance. We identify several disadvantages of using dense captions as visual information, which could potentially lead to a decline in performance. (1) They usually lack a high-level understanding of the image content, which is essential to answer many questions. (2) Although they provide a caption for each object, these captions are general and still lack the key information about an object to answer the question, e.g. action of a human, or the color of the cloth. (3) Since dense captions are focused solely on objects, they neglect broader contextual information such as weather conditions. (4) The captions for most objects are not useful and could mislead or confuse LLMs due to overwhelming information. This further shows that acquiring important visual information based on the question context is effective.

Table 4: Ablations on VCR with more decoding configurations.

Decision Model	**Decoding Config**	**VCU**		**VCI**	
		Conf	!Conf	Conf	!Conf
LLM + Caption + LLM Clue	Orig	72.9	58.1	57.1	52.9
	Temp 0.1	72.4	58.8	57.1	53.9
	Temp 0.2	72.4	58.5	61.2	52.2
	ICL examples	74.7	56.7	59.2	54.2
Num. of Examples		170	1779	49	1002

Table 5: The result of ViCor on open-ended VCR datasets.

Method	OKVQA	AOKVQA
LLM+Caption	34.6	44.1
BLIP2-T5XL	36.2	49.7
ViCor (ours)	**38.7**	**50.9**

Table 6: The result of ViCor on OKVQA dataset.

Method	Accuracy
LLM+Caption	63.3
LLM+Dense Caption	40.5
ViCor (ours)	**70.9**