Enhancing Advanced Visual Reasoning Ability of Large Language Models

Anonymous ACL submission

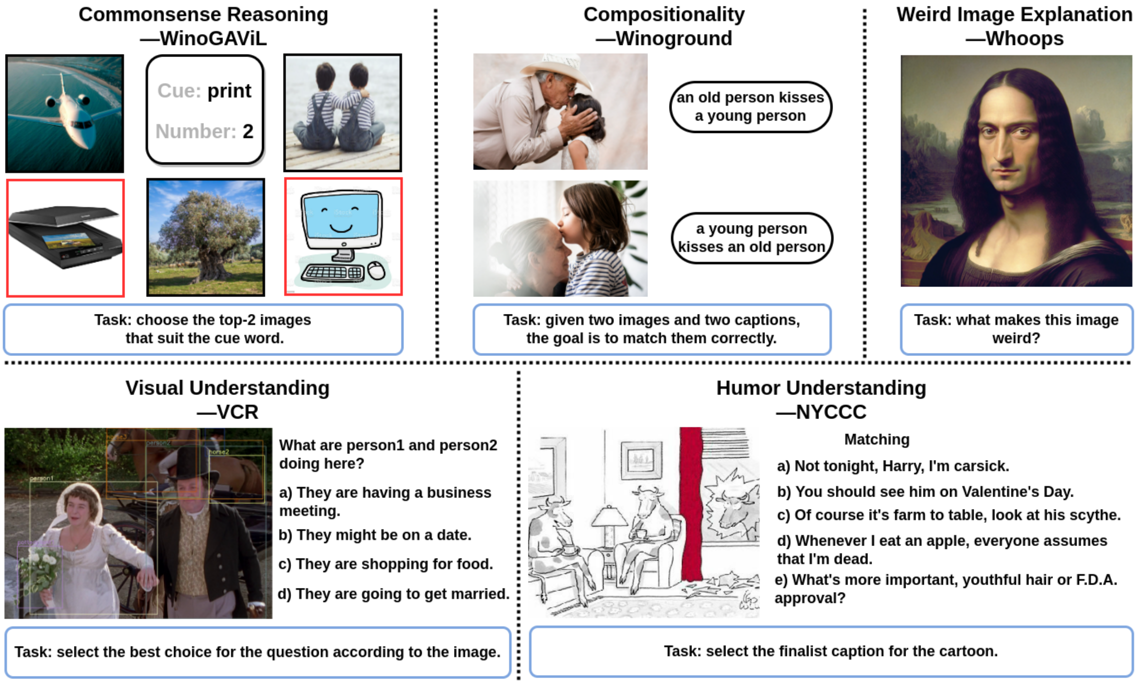
Abstract

Recent advancements in Vision-Language (VL) research have sparked new benchmarks for complex visual reasoning, challenging models' advanced reasoning ability. Traditional Vision-Language models (VLMs) perform well in visual perception tasks while struggling with complex reasoning scenarios. Conversely, Large Language Models (LLMs) demonstrate robust text reasoning capabilities; however, they lack visual acuity. To bridge this gap, we propose **C**omplex **V**isual **R**easoning **L**arge **L**anguage **M**odels (CVR-LLM), capitalizing on VLMs' visual perception proficiency and LLMs' extensive reasoning capability. Unlike recent multimodal large language models (MLLMs) that require a projection layer, our approach transforms images into detailed, context-aware descriptions using an iterative self-refinement loop and leverages LLMs' text knowledge for accurate predictions without extra training. We also introduce a novel multimodal in-context learning (ICL) methodology to enhance LLMs' contextual understanding and reasoning. Additionally, we introduce Chain-of-Comparison (CoC), a step-by-step comparison technique enabling contrasting various aspects of predictions. Our CVR-LLM presents the first comprehensive study across a wide array of complex visual reasoning tasks and achieves SOTA performance among all.

1 Introduction

The concept of complex visual reasoning was introduced with Visual Commonsense Reasoning (VCR) dataset (Zellers et al., 2019) in 2019, which tests models' ability to understand visual content as well as commonsense cognition. However, the development in this field has remained relatively subdued, primarily due to Vision-and-Language Models' (VLMs) limitations in incorporating commonsense knowledge (Gan et al., 2022). Recent years have seen significant advancements in complex linguistic reasoning tasks (Cobbe et al., 2021; Wei et al., 2022) due to the emerging GPT3 (Brown et al., 2020), LLaMA (Touvron et al., 2023a), and Vicunna (Chiang et al., 2023). This leap forward has triggered a renewed interest in the complex visual reasoning area, exploring how visual perception can enhance linguistic inference and potentially overcome previous hurdles (Gan et al., 2022). It has led to innovative benchmarks focusing on various aspects: commonsense reasoning - WinoGAViL (Bitton et al., 2022), compositionality - Winoground (Thrush et al., 2022), weird image explanation - Whoops (Bitton-Guetta et al., 2023), and humor understanding - NYCCC (Hessel et al., 2022). These tasks demand models not only accurately interpret image content, but also integrate knowledge from daily experiences, general commonsense, cultural context, and humor sense. For example, a synthetic image, as shown in Whoop's example in Figure 1 of "The portrait of the Mona Lisa depicts a stern male face." contradicts the cultural context, as the famous painting Mona Lisa depicts a female face.

In this paper, we introduce a novel method named Complex Visual Reasoning Large Language Models (CVR-LLM), based on the "VLMs + LLMs" concept. Recent Multimodal large language models (MLLMs) like LLaVA (Liu et al., 2024, 2023a) and MiniGPT4 (Zhu et al., 2023; Chen et al., 2023) have proven effective in many VL tasks. However, these models are resource-intensive, relying on millions of image-text pairs for projection layer learning. To overcome this limitation, our approach leverages the visual perception strengths of VLMs to translate images into context-aware image descriptions (CaID) via an inference-only, dual-loop self-refinement process that incorporates feedback from LLMs. These detailed descriptions enhance the LLMs' inference process, transforming multi-modal tasks into simpler single-modal challenges and streamlining the overall process. In addition, we develop a unique



Task: choose the top-2 images that suit the cue word.

Task: given two images and two captions, the goal is to match them correctly.

Task: what makes this image weird?

What are person1 and person2 doing here?
a) They are having a business meeting.
b) They might be on a date.
c) They are shopping for food.
d) They are going to get married.

Task: select the best choice for the question according to the image.

a) Not tonight, Harry, I'm carsick.
b) You should see him on Valentine's Day.
c) Of course it's farm to table, look at his scythe.
d) Whenever I eat an apple, everyone assumes that I'm dead.
e) What's more important, youthful hair or F.D.A. approval?

Task: select the finalist caption for the cartoon.

Figure 1: Five distinct examples from diverse datasets in the complex visual reasoning field (Bitton-Guetta et al., 2023) challenge AI models' ability of complex reasoning in different aspects such as general commonsense.

multi-modal ICL approach named Complex Visual Reasoning ICL (CVR-ICL), which enhances the reasoning capacities of LLMs within a range of complex multi-modal environments. Figure 2 provides an illustration of how our CVR-LLM is applied to the Winoground task. It describes the images as appropriate sentences via CaID and utilizes the sophisticated reasoning and ICL abilities of LLMs through CVR-ICL for more accurate predictions.

Our research stands as the pioneering study to explore such a broad array of benchmarks (WinoGAViL, Winoground, Whoops, VCR, and NYCCC), proposing a paradigm centred on the "VLM+LLM" concept for addressing complex visual reasoning tasks. Experimental results show that CVR-LLM achieves SOTA performance across all five tasks. Further ablation studies and comparative analyses reveal the effectiveness of each module and the superiority of our method over previous approaches. Particularly in comparative analysis, we introduce the Chain-of-Comparison (CoC) technique, inspired by "Chain-of-Thought" and utilizing GPT4 (Achiam et al., 2023), to address the limitations of conventional metrics in evaluating abstract concepts. CoC provides a nuanced analysis by systematically dissecting and quantitatively contrasting various facets of the results for a comprehensive evaluation.

Our contributions are summarized as follows: (1) We present the first comprehensive study across all complex visual reasoning tasks, including WinoGAViL, Winoground, Whoops, VCR, and NYCCC. (2) We design a context-aware image description generation method and a specific in-context learning strategy, to enhance the advanced visual reasoning ability of LLMs to multi-modal complex visual reasoning tasks. (3) We further introduce Chain-of-Comparsion, a novel GPT4-based comparison technique inspired by "Chain-of-Thought" filling the gaps of traditional metrics in abstract concept evaluation. (4) Experimental results show that our approach surpasses current SOTA models in a range of complex visual reasoning scenarios.

2 Related Work

2.1 Reasoning Research in Vision-Language Domain

In recent years, multi-modal reasoning research has significantly advanced. Beyond the complex visual reasoning benchmarks discussed in Section 1, many studies focus on the reasoning process itself, such as chain-of-thought (Kojima et al., 2022; Shaikh et al., 2022) or reasoning modules (Zhou et al., 2023b; Jiang et al., 2023), which are crucial for enhancing AI models' analytical capabilities and performance. For instance, Liu et al. (2023b) introduced a modality-aligned thought chain reasoning framework to incorporate explicit reasoning into task-oriented dialogue generation, improving contextual understanding and effectiveness.

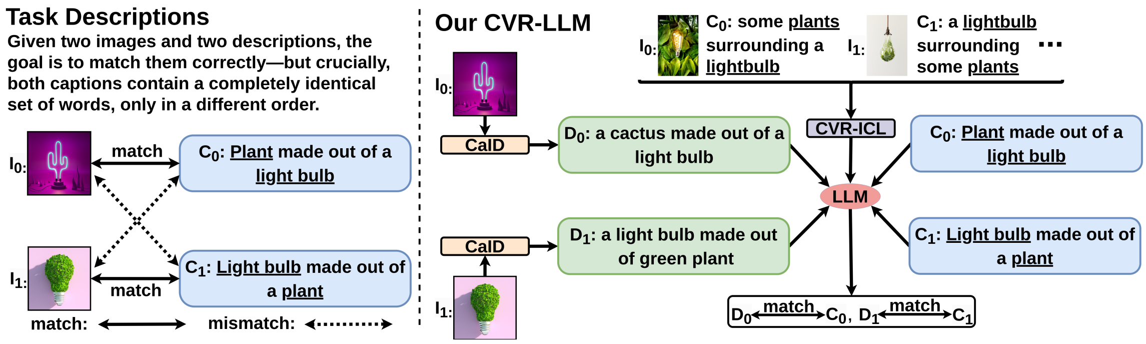

Task Descriptions
Given two images and two descriptions, the goal is to match them correctly—but crucially, both captions contain a completely identical set of words, only in a different order.
I_0:
match
C_0: Plant made out of a light bulb
I_1:
match
C_1: Light bulb made out of a plant
match:
mismatch:
Our CVR-LLM
I_0:
CaID
D_0: a cactus made out of a light bulb
CaID
D_1: a light bulb made out of green plant
I_1:
I_0:
C_0: some plants surrounding a lightbulb
I_1:
C_1: a lightbulb surrounding some plants
...
CVR-ICL
C_0: Plant made out of a light bulb
LLM
C_1: Light bulb made out of a plant
$D_0 \xleftrightarrow{match} C_0$, $D_1 \xleftrightarrow{match} C_1$

Figure 2: An example of our CVR-LLM works on the Winoground dataset. Our method transfers images into context-aware image descriptions through CaID and leverages the sophisticated reasoning and ICL abilities of LLMs with the CVR-ICL module, offering a more precise answer.

Lv et al. (2023) proposed a counterfactual cross-modality reasoning method for better video moment localization. Zhou et al. (2023a) developed a multi-step reasoning probability transfer mechanism to improve multi-label interaction classifications. Yu et al. (2023) presented a hierarchical reasoning network to consolidate multi-level interactive cues, from coarse to fine-grained details, enhancing Human-Object Interaction (HOI) representations.

2.2 Large Language Models for Vision-Language Analysis

The past two years have seen an unprecedented surge in the development and application of LLMs (Brown et al., 2020; Touvron et al., 2023a; Chiang et al., 2023) across diverse fields. LLMs have garnered acclaim for their robust capabilities, including advanced analytical prowess (Kojima et al., 2022), extensive text-level knowledge (Naveed et al., 2023) and superior understanding ability (Chang et al., 2023). Furthermore, they are equipped with two powerful mechanisms: chain-of-thought (Kojima et al., 2022) and in-context learning (Liu et al., 2021a), which significantly augment their effectiveness and performance in specialized tasks (Naveed et al., 2023). For example, Muraoka et al. (2023) developed a cross-lingual model trained alongside a cross-lingual LLM, leveraging LLMs' capabilities across languages. Lan et al. (2023) proposed reasoning question prompts for Visual Question Answering (VQA) tasks, unlocking LLMs' potential in zero-shot learning. Additionally, Yang et al. (2023) introduced SODA, a system that integrates LLMs with explainable AI to assist marketers with data interpretation, enhancing human-AI collaboration. Zhong et al. (2023) used knowledge distillation to imbue the SUR-adapter with LLMs' semantic understanding and reasoning capabilities.

3 Methods

In this section, we introduce the CVR-LLM framework, highlighting its innovative process for generating context-aware image descriptions (CaID) as well as its complex visual reasoning in-context learning (CVR-ICL) strategy. Initially, we explain the CaID generation process, which differs from traditional image captioning by using a self-refinement loop with feedback from Large Language Models (LLMs) to produce accurate and contextually relevant descriptions (Section 3.1). Subsequently, we present the CVR-ICL approach (Section 3.2), which enhances LLMs' contextual understanding and reasoning by assessing relevant cases and selecting suitable complex multi-modal demonstrations.

3.1 Context-Aware Image Description

Pre-trained VLMs (Li et al., 2023; Alayrac et al., 2022) have demonstrated their proficiency in generating detailed image captions on benchmarks such as MSCOCO (Chen et al., 2015). However, while these captions may accurately reflect visual content, they are not customized for complex visual reasoning scenarios. Recently, the trend of multi-modal instruction-following agents like miniGPT4 (Zhu et al., 2023; Chen et al., 2023) and LLaVA (Liu et al., 2024, 2023a), integrating open-source LLMs (Chiang et al., 2023; Touvron et al., 2023b) with pre-trained vision encoders (Dosovitskiy et al., 2020; Liu et al., 2021b) to create a MLLM, has become very popular. The effectiveness of these models is heavily reliant on tuning with vast amounts of VL instruction data, which is generated by powerful LLMs like ChatGPT (OpenAI, 2022) and GPT4 (Achiam et al., 2023). While promising, their reliance on extensive VL instruction data for tuning requires the substantial resource and time investment. In this work, we introduce a

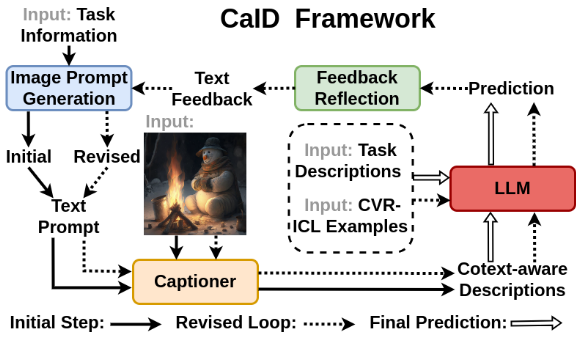


Figure 3: The framework overview of CaID. It is designed to transfer images into contextualized descriptions, bypassing the need for direct multi-modal fusion and leveraging LLMs' extensive knowledge for more accurate predictions.

more efficient method for generating context-aware image descriptions, which depends on the inference process and leverages task-specific information and feedback from LLMs to craft better prompts, guiding the caption generation process more effectively.

Our CaID framework optimizes the process of creating context-aware image descriptions through a dual-loop self-refinement approach, as shown in Figure 3. Initially, it leverages task-specific details and LLM insights to craft precise image prompts. These initial prompts are designed to distill essential task-related information, guiding the captioner in producing descriptions that are not only cover image content but also deeply aligned with the task's requirements. Specifically, given a task specific text description t with an image i (for processes involving multiple images, we approach each image sequentially), the generation of initial context-aware image descriptions can be described as follows:

$$d_{init} = C(i, L(t)), \tag{1}$$

where d_{init} is the initial generated context-aware image description. C is the image-to-text captioner, transfering the image into the description. L is the LLM, encapsulating crucial task-related text information t (e.g. requirements, questions, cue words) into feature prompts.

In the second loop, our approach is crafted to encapsulate essential task-related details as well as LLMs' feedback, enhancing description generation with LLMs' vast knowledge. Specifically, it merges initial descriptions with task specifics and CVR-ICL examples into a task-focused prompt, guiding LLMs to make more precise predictions. These predictions are then treated as pseudo labels, asking LLMs to design further inquiries for deeper insights around them. In this way, we build up a feedback reflection between LLM prediction and context-aware caption, enhancing the richness and accuracy of the content produced. The textual feedback is then leveraged to refine the image prompts, providing deep insights that inform and guide the generation of nuanced image descriptions. The revised context-aware image descriptions can be described as follows:

$$d_{revised} = C(i, L(t, Q(p))), \tag{2}$$

where $d_{revised}$ is the revised generated context-aware image description. Q is the further query from LLM. p is the prediction from LLM according to the generated task prompt. $Q(p)$ is the text feedback for updating image prompt.

3.2 Complex Visual Reasoning ICL

LLMs are renowned for their exceptional in-context learning capabilities, especially with task-specific examples. The optimal in-context exemplars enable LLMs to leverage their background knowledge for more precise outcomes. However, most of the research works (Liu et al., 2021a; Sorensen et al., 2022) have primarily focused on the text-centric domain, with few works (Alayrac et al., 2022; Zhao et al., 2023) exploring multi-modal in-context learning for VL tasks. Our approach, unlike prior methods focused solely on text similarity in NLP, such as the kNN-augmented in-context example selection (KATE), integrates multi-modal factors, thereby enriching the discipline with a fresh perspective. Furthermore, it is also different from MMICL (Zhao et al., 2023) in the multi-modal domain, which employs a vision prompt generator for image-to-visual embedding conversion and merges these with text embeddings as a union measurement factor.

Complex visual reasoning tasks demand models capable of selecting in-context examples from a multi-modal domain, leveraging extensive background knowledge and information within it (Zhao et al., 2023). However, our CVR-LLM is grounded in LLMs, which are inherently text-based, leading to a gap between textual and multi-modal domains. Directly applying a text-based kNN clustering method could result in the loss of important multi-modal information. On the other hand, using multi-modal information for retrieval might ignore essential context-aware information within our generated image descriptions. To address this, we

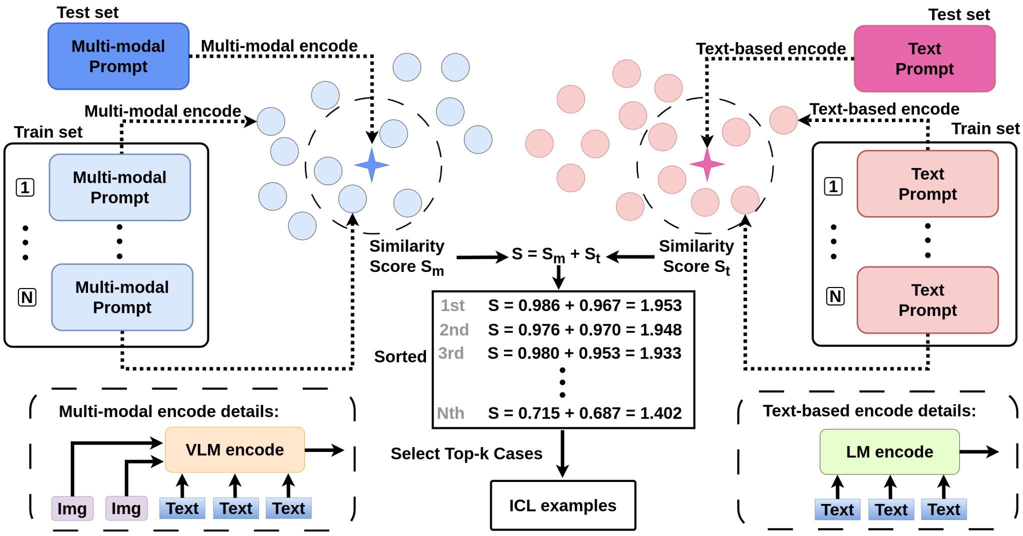


Figure 4: The generic diagram of our proposed CVR-ICL approach. The dual analysis enables our approach to more effectively select contextually relevant examples from text and multi-modal domains.

propose the complex visual reasoning ICL, which aims to select in-context examples for LLMs by effectively integrating both text and multi-modal components. This dual analysis enables our LLM to more effectively select contextually relevant examples. ensuring a balanced integration of text and multi-modal insights for enhanced in-context learning. Figure 4 illustrates the framework of our CVR-ICL strategy. Specifically, given a task t with an image i, we initially convert the image into a description d, which enables the task to be applicable not only in multi-modal domains but also in text-only scenarios. Then, we employ a multi-modal encoder f_m and a text encoder f_t to transform inputs from the multi-modal domain and the text domain into vector representations as follows:

$$x_m = f_m(t, i), \tag{3a}$$

$$x_t = f_t(t, d), \tag{3b}$$

where x_m is the vector representation in the multi-modal domain. x_t is the vector representation in the text domain.

Upon transforming each example into two distinct vector forms, we compute the cosine similarity score to identify and select the examples that are most relevant. Considering a target sample in test set and the ith example in the training set, the similarity calculation process can be expressed as follows:

$$s_m = f_c(x_m, x_m^{ith}), \tag{4a}$$

$$s_t = f_c(x_t, x_t^{ith}), \tag{4b}$$

$$s = s_m + s_t, \tag{4c}$$

where s_m is the similarity score between the target sample and ith example in dataset on the multi-modal domain, s_t is the similarity score between the target sample and ith example in dataset on the text domain. s is the final similarity score. f_c is the cosine similarity function. Finally, the top-k cases with the highest s are selected as the in-context examples, aimed at boosting the contextual understanding and prediction accuracy of the LLMs.

4 Experiments

4.1 Dataset and Metrics

To evaluate the effectiveness of our proposed method, we conduct a comprehensive test in complex visual reasoning areas. Our evaluation included WinoGAViL (4373 samples), Winoground (400 samples), Whoops (500 samples), VCR (2653 out of over 26k samples, selecting a random 10%), and NYCCC (528 samples), providing a broad assessment of our approach's capabilities. In the terms of metrics, we adhered to the evaluation methods provided by these datasets, ensuring a fair assessment of our method's performance.

4.2 Implementation Details

For the basic captioner in context-aware image description (Section 3.1), we choose the BLIP2-flant5xxl (Li et al., 2023) as our baseline. For CVR-ICL phase (Section 3.2), we employ BM25 (Robertson et al., 1995) and BLIP2 multi-embedding (Li et al., 2023) to encode text and multi-modal inputs, respectively. It's important to note that the ICL example results are derived from LLM inference without using actual annotations to prevent data leakage. For our LLMs, we choose three popular LLMs as inference models for generation

Type	Model	WinoGAViL			Winoground			Whoops	VCR		NYCCC	
		5/6	10/12	SWOW	Text	Image	Group	GPT4 Rate	Q->A	QA->R	Match acc.	CrowdAcc
VLM	ViLT (2021)	55.0	52.0	59.0	34.7	14.0	9.2	-	-	-	-	-
	CLIP ViT-L/14 (2021)	47.0	15.0	66.0	-	-	-	-	-	-	56.6	55.8
	UNITER (2020)	-	-	-	38.0	14.0	10.5	-	-	-	-	-
	ViLLA (2020)	-	-	-	37.0	13.2	11.0	-	-	-	48.1	47.0
	BLIP (2022)	54.6	45.0	66.5	**46.5**	27.7	24.2	22.0	29.2	27.5	58.7	58.1
	BLIP2 (2023)	49.3	38.8	71.6	44.0	26.0	23.5	31.0	24.5	25.6	58.3	56.7
MLLM	LLaVA 1.0 (2024)	-	-	-	-	-	-	32.0	28.3	40.0	55.8	53.1
	LLaVA 1.5 (2023a)	-	-	-	-	-	-	42.4	35.1	44.5	59.3	56.0
	MiniGPT4 V1 (2023)	-	-	-	-	-	-	44.6	40.6	47.7	58.5	55.6
	MiniGPT4 V2 (2023)	-	-	-	-	-	-	48.2	48.8	49.7	60.4	**59.2**
VLM+LLM	CVR-LLM$_{Llama3}$	72.3	70.4	**88.7**	45.0	29.5	24.5	60.4	50.5	52.4	59.8	57.7
	CVR-LLM$_{GPT3.5}$	73.4	71.6	83.4	42.7	30.5	23.5	61.2	51.1	53.4	59.4	56.8
	CVR-LLM$_{GPT4}$	**74.7**	**73.2**	86.5	43.5	**35.0**	**26.5**	**62.0**	**52.9**	**54.3**	**60.6**	57.4

Table 1: The comparison of our CVR-LLM with popular VLMs and MM LLMs on five complex visual reasoning tasks. Notably, MLLMs like LLaVA and MiniGPT4 exhibit limitations in handling tasks involving multiple images or computing image-text similarity scores, resulting in their performance being unavailable for tasks like WinoGAViL and Winoground.

tests including: Llama3-8B (Meta, 2024) for CVR-LLM$_{Llama3}$, GPT3.5 (OpenAI, 2023) for CVR-LLM$_{GPT3.5}$, and GPT4 (Achiam et al., 2023) for CVR-LLM$_{GPT4}$. Performance comparisons are conducted directly on the test set without any fine-tuning, as WinoGAViL, Winoground, and NYCC datasets are exclusively for testing purposes.

4.3 Comparison to State-of-the-Arts

In this section, we evaluate our proposed CVR-LLM against various models across a range of complex visual reasoning tasks, including WinoGAViL, Winoground, Whoops, VCR, and NYCCC. These models fall into two categories: VLMs (Kim et al., 2021; Radford et al., 2021; Gan et al., 2020; Li et al., 2023) and MLLMs (Liu et al., 2024, 2023a; Zhu et al., 2023; Chen et al., 2023). Notably, MLLMs like LLaVA and MiniGPT4 struggle with tasks involving multiple images, making their performance data unavailable for WinoGAViL and Winoground.

Table 1 showcases our method's superiority across five tasks, eclipsing both VLMs and LMMs. For example, our CVR-LLM$_{Llama}3$ significantly surpasses the SOTA model BLIP2 by achieving an 88.7% accuracy (+17.1 improvement) in SWOW setting on the WinoGAViL benchmarks. Similarly, it outperforms the SOTA model MiniGPT4 with a 62.0% accuracy (+13.8 improvement) on the GPT4 rate (Bitton-Guetta et al., 2023) for Whoops tasks, underscoring our framework's advanced performance. Additionally, our method performs well on three LLM-based categories, demonstrating robust generation abilities with consistent performance. This highlights the versatility and adaptability of our model, ensuring high-quality results across various complex visual reasoning tasks.

4.4 Ablation Studies

In this section, we examine the individual contributions of the components within our framework CVR-LLM$_{GPT4}$. As demonstrated in Table 2, we present an ablation study that quantifies the performance impact of each module across various datasets. The experimental findings suggest that the CVR-ICL module significantly boosts the inference performance of LLMs compared to using context-aware image descriptions alone, with the exception of the NYCCC dataset (It may be due to NYCCC's focus on humor, where precise descriptions are more critical). This highlights the CVR-ICL module's effectiveness in enhancing LLM capabilities across various tasks. In addition, our comprehensive method, CVR-LLM, which integrates both context-aware descriptions and CVR-ICL, achieves a substantial enhancement in performance relative to the baseline.

4.5 Analysis

Context-Aware Image Description vs General Image Caption In this section, we investigate CaID's impact at an abstract level and design a novel method to quantitatively demonstrate the semantic gap between context-aware image descriptions and general image captions (Note that the performance impact has been shown in Table 2). Figure 5 provides two examples comparing context-aware image descriptions with general image captions and our goal is to determine whether context-aware descriptions offer more contextually relevant information to aid LLMs in decision-making. Unlike traditional sentence evaluations that rely on

Module	WinoGAViL			Winoground			Whoops	VCR			NYCCC		
	5/6	10/12	SWOW	Text	Image	Group	GPT4 Rate	Q->A	QA->R	Q->AR	Match acc.	CrowdAcc	NYAcc
Base	60.0	58.3	78.4	28.7	26.2	16.0	36.4	38.0	37.0	21.3	41.8	41.3	46.0
Base+CaID	63.5	62.0	73.7	31.5	30.0	19.7	54.6	43.9	44.2	22.9	51.5	48.7	53.6
Base+CVR-ICL	69.8	66.1	80.9	39.0	29.2	22.0	60.6	48.8	49.2	25.8	48.0	47.6	52.9
CVR-LLM$_{GPT4}$	**73.4**	**73.2**	**86.5**	**43.5**	**35.0**	**26.5**	**62.0**	**54.3**	**52.9**	**30.4**	**60.6**	**57.4**	**63.1**

Table 2: The ablation study of our CVR-LLM on five complex visual reasoning tasks. "Base" represents using the general image captions and GPT4 to complete these tasks. "Base+CaID" means using the context-aware image descriptions instead of the general image captions and GPT4 to test the performance. "Base+CVR-ICL" represents using general image captions and GPT4 with our designed CVR-ICL learning methods.

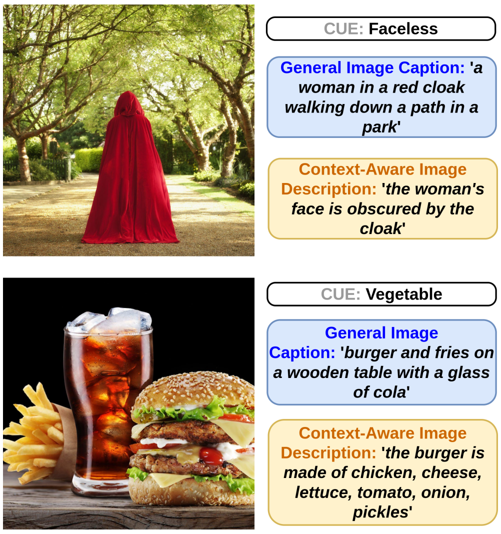
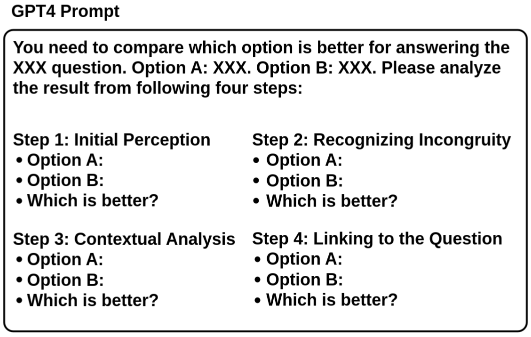

CUE: Faceless
General Image Caption: 'a woman in a red cloak walking down a path in a park'
Context-Aware Image Description: 'the woman's face is obscured by the cloak'
CUE: Vegetable
General Image Caption: 'burger and fries on a wooden table with a glass of cola'
Context-Aware Image Description: 'the burger is made of chicken, cheese, lettuce, tomato, onion, pickles'

Figure 5: Two examples from WinoGAViL compare context-aware image descriptions with general image captions. WinoGAViL is designed to ask the model to select the image that best matches the cue word.



GPT4 Prompt

You need to compare which option is better for answering the XXX question. Option A: XXX. Option B: XXX. Please analyze the result from following four steps:

Step 1: Initial Perception
- Option A:
- Option B:
- Which is better?

Step 2: Recognizing Incongruity
- Option A:
- Option B:
- Which is better?

Step 3: Contextual Analysis
- Option A:
- Option B:
- Which is better?

Step 4: Linking to the Question
- Option A:
- Option B:
- Which is better?

Figure 6: The illustration of how to use GPT4 for step-by-step comparsion.

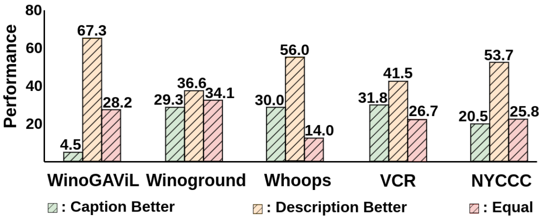


Figure 7: Hypothesis verification with GPT4, which demonstrates the effectiveness of our CaID against general image captions.

annotations to compute metrics like BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), we lack direct measures to assess the contextual relevance of sentences. To address this, we use GPT4 (Achiam et al., 2023) to evaluate the relative effectiveness between two kinds of expressions with the prompt: "*Evaluate the equivalence of the following two options for the task XXX. Option A: XXX; Option B: XXX. Please return True if Option B is better than Option A in answering questions; return False if the opposite is true; return Equal if they are the same for the question.*". Additionally, inspired by the concept of chain-of-thought (CoT) (Wei et al., 2022), we propose a novel comparison chain-of-comparison (CoC), which implements a step-by-step analysis to evaluate the effectiveness. This method involves a comprehensive four-step analysis protocol, depicted in Figure 6. It follows a series of cognitive steps that our brains undertake to make sense of information, particularly when engaging with complex problems.

Figure 7 shows the results of directly employing GPT4 to compare the effectiveness of general image captions with our image descriptions in the specific scenario of answering task-related questions. Furthermore, Table 3 presents the performance derived from utilizing GPT4 to conduct a detailed, step-by-step analytical assessment of effectiveness. These empirical results indicate that our approach yields image descriptions with enhanced contextual relevance, thereby significantly aiding LLMs in the decision-making process, particularly on the WinoGAViL and Whoops datasets.

Complex Visual Reasoning ICL vs Other ICL

The CVR-ICL is designed to optimize the selection of in-context exemplars within a multi-modal environment, thereby enhancing the reasoning abilities of LLMs. This innovative method is contrasted with three alternative configurations: Random In-Context Learning (RICL) (Brown et al., 2020), KATE (Liu et al., 2021a), and Multi-modal Similar In-Context Learning (MMICL) (Zhao et al., 2023). To ensure a fair comparison, we utilized general image captions across all models to test performance for eliminating the effect of our context-aware im-

Dataset	Option	Step 1	Step 2	Step 3	Step 4	Average
WinoGAViL	Caption Better	6.0	4.3	8.3	5.0	5.9
	Description Better	75.3	76.0	71.3	76.7	**74.8**
	Equal	18.7	19.7	20.3	18.3	19.3
Winoground	Caption Better	24.0	24.0	29.0	27.0	26
	Description Better	59.0	56.0	59.0	56.0	**57.5**
	Equal	17.0	20.0	12.0	17.0	16.5
Whoops	Caption Better	27.0	13.0	14.0	13.0	16.7
	Description Better	71.0	80.0	76.0	75.0	**75.5**
	Equal	2.0	7.0	10.0	12.0	7.7
VCR	Caption Better	24.3	32.5	30.1	28.6	28.9
	Description Better	53.5	45.4	50.6	52.7	**50.5**
	Equal	22.2	22.1	19.3	18.7	20.6
NYCCC	Caption Better	18.6	15.8	17.4	19.1	17.7
	Description Better	58.5	62.3	60.4	61.0	**60.5**
	Equal	22.9	21.9	22.2	19.9	21.8

Table 3: The performance of using GPT4 to assess the effectiveness of two options (general image caption and our context-aware image description) based on CoC.

Dataset	Category	RICL (2020)	KATE (2021a)	MMICL (2023)	CVR-ICL
WinoGAViL	5/6	64.1	68.6	66.3	**69.8**
	10/12	61.7	64.1	62.8	**66.1**
	SWOW	80.7	**82.8**	80.9	80.9
Winoground	Text	35.0	29.5	27.5	**39.0**
	Image	22.5	30.0	25.0	**29.2**
	Group	18.5	**20.0**	17.5	22.0
Whoops	GPT4 Rate	60.4	62.0	60.8	**62.0**
VCR	Q->A	45.1	48.6	44.0	**48.8**
	QA->R	46.5	48.9	46.3	**49.2**
	Q->AR	22.5	24.8	23.6	**25.8**
NYCCC	Match acc.	44.4	47.5	45.5	**48.0**
	CrowdAcc	46.6	46.4	43.7	**47.6**
	NYAcc	50.3	51.2	49.8	**52.9**

Table 4: The performance of using different ICL methods on different datasets.

age descriptions. As demonstrated in Table 4, our CVR-ICL outperforms other ICL methods, demonstrating its adeptness at integrating and leveraging both textual and multi-modal domains to select the most contextually appropriate exemplars.

Case Number Selection in Complex Visual Reasoning ICL Figure 8 illustrates the influence of varying case numbers in the CVR-ICL on the performance of our proposed CVR-LLM method. The experimental results suggest a trend where the model's performance initially improves with an increase in case numbers, exhibits fluctuations at higher numbers, and eventually declines as the case number becomes excessively large. This pattern suggests that the optimal selection for the number of cases is four.

5 Qualitative Results

To showcase our approach capabilities, we present qualitative results in Figure 9. It illustrates how LLMs leverage contextual information to ask more relevant and insightful questions tailored the specific tasks. For instance, when provided with an image of the chess piece, the LLMs might ask "What does the chess piece look like?". Subsequently, the captioner model generates contextually appropriate descriptions, such as "A chess piece that looks like a unicorn.". This synergy enhances the LLM's decision-making process, making it more precise and context-aware. More detailed qualitative results with corresponding prompts and CVR-ICL examples are illustrated in Appendix A.1 and Appendix A.2.

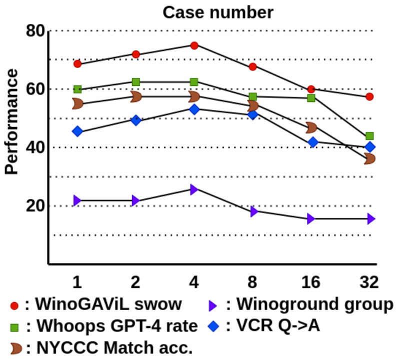


Figure 8: The different case numbers in CVR-ICL and corresponding performance.

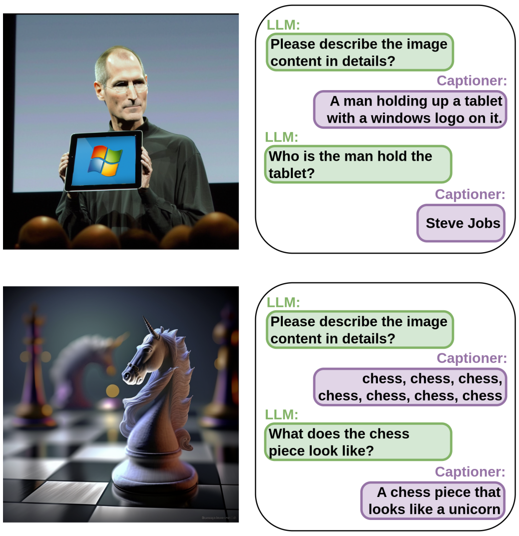

LLM:
Please describe the image content in details?
Captioner:
A man holding up a tablet with a windows logo on it.
LLM:
Who is the man hold the tablet?
Captioner:
Steve Jobs
LLM:
Please describe the image content in details?
Captioner:
chess, chess, chess, chess, chess, chess, chess
LLM:
What does the chess piece look like?
Captioner:
A chess piece that looks like a unicorn

Figure 9: Two qualitative results from Whoops illustrating the capabilities of our approach. Whoops is designed to ask the model to explain what makes images weird.

6 Conclusion

In this work, we propose CVR-LLM, an innovative approach for complex visual reasoning tasks. This method boosts LLMs' understanding of visual content for complex reasoning via context-aware image descriptions. We also develop a multi-modal in-context learning technique, enhancing LLMs' reasoning skills at both image and text levels. Experimental results show that CVR-LLM sets new benchmarks across multiple complex visual reasoning tasks. We also introduce a nuanced GPT4 based analysis technique Chain-of-Comparison to automatically break down and contrast among various aspects of generated results.

7 Limitation

Although our approach achieves SOTA performance across a wide range of complex visual reasoning benchmarks, it still has two notable limitations. First, compared to the MLLMs that can perform end-to-end inference directly, our approach operates as an LLM-agent-driven framework. This involves VLMs generating context-aware image descriptions, followed by the LLM performing inference with ICL to predict the answer. While this two-step process enhances contextual understanding and reasoning, it may significantly increase time consumption compared to direct end-to-end inference models. Second, despite its overall strong performance and generalization ability, our approach still lags behind GPT4V in some tasks. Figure 10 shows that our CVR-LLM can surpass GPT4V in SWOW setting in WinoGAViL dataset but fall short in others. Our future work will focus on refining the integration between VLMs and LLMs components and enhancing the model's efficiency and accuracy across a broader spectrum of complex visual reasoning challenges.

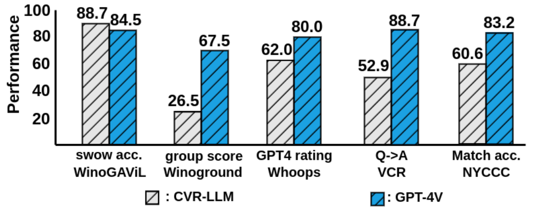


Figure 10: The comparison of our CVR-LLM against GPT-4V.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. 2022. Winogavil: Gamified association benchmark to challenge vision-and-language models. *Advances in Neural Information Processing Systems*, 35:26549–26564.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. In *arXiv preprint arXiv:1504.00325*.

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 *European conference on computer vision*, pages 104–120. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pretraining: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. 2023. Multi-spectral image stitching via spatial graph reasoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 472–480.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

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.

Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4389–4400.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yiting Liu, Liang Li, Beichen Zhang, Shan Huang, Zheng-Jun Zha, and Qingming Huang. 2023b. Matcr: Modality-aligned thought chain reasoning for multimodal task-oriented dialogue generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5776–5785.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zezhong Lv, Bing Su, and Ji-Rong Wen. 2023. Counterfactual cross-modality reasoning for weakly supervised video moment localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6539–6547.

Meta. 2024. Llama 3 model card.

Masayasu Muraoka, Bishwaranjan Bhattacharjee, Michele Merler, Graeme Blackwood, Yulong Li, and Yang Zhao. 2023. Cross-lingual transfer of large language model by visually-derived supervision toward low-resource languages. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3637–3646.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-3.5: Generative pre-trained transformer 3.5. `https://www.openai.com/research/gpt-3-5`. Accessed: 2023-06-12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575. IEEE.

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.

Qi Yang, Marlo Ongpin, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. 2023. Against opacity: Explainable ai and large language models for effective digital advertising. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9299–9305.

Jiale Yu, Baopeng Zhang, Qirui Li, Haoyang Chen, and Zhu Teng. 2023. Hierarchical reasoning network with contrastive learning for few-shot human-object interaction recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4260–4268.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. 2023. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578.

Yuchen Zhou, Guang Tan, Mengtang Li, and Chao Gou. 2023a. Learning from easy to hard pairs: Multi-step reasoning network for human-object interaction detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4368–4377.

Zhuo Zhou, Wenxuan Liu, Danni Xu, Zheng Wang, and Jian Zhao. 2023b. Uncovering the unseen: Discover hidden intentions by micro-behavior graph reasoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6623–6633.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

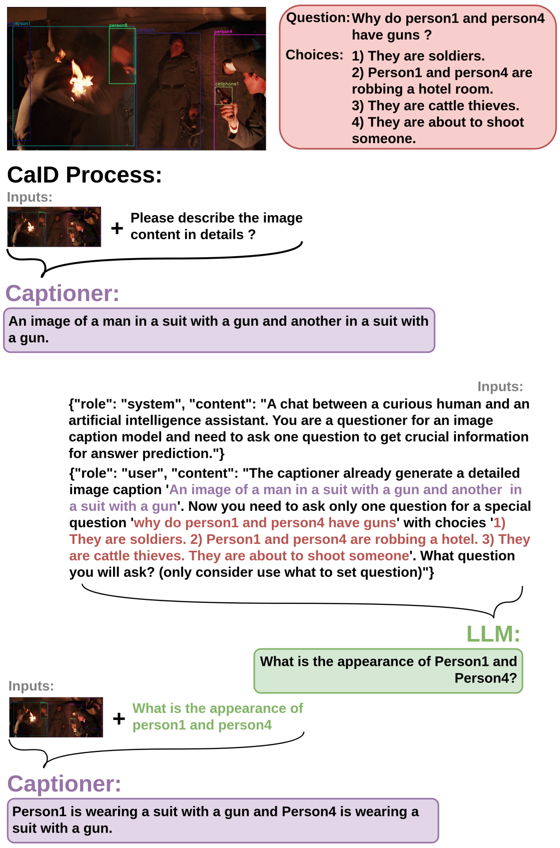
A.1 Qualitative Results with Corresponding Prompt

Section 5 only illustrates the simplified process of our Context-aware Image Description (CaID) generation. Here, we delve into more details about the generation process and the corresponding prompts. Figure 11 provides an example of the CaID generation process applied to the VCR (Zellers et al., 2019) task. In this example, the initial input consists of an image showing several individuals, with two of them (Person1 and Person4) holding guns. The associated question is: "Why do Person1 and Person4 have guns?" with multiple-choice options such as "1) They are soldiers. 2) Person1 and Person4 are robbing a hotel room. 3) They are cattle thieves. 4) They are about to shoot someone.".

The CaID process begins by generating a detailed description of the image. The captioner model produces an initial caption: "An image of a man in a suit with a gun and another in a suit with a gun." This caption, while descriptive, lacks the context needed to answer the specific question posed. To address this, our system prompts the LLM with a scenario where it acts as a questioner for the image caption model. The LLM is instructed to generate a follow-up question to gather crucial information for answer prediction. The prompt guides the LLM to consider specific details such as the appearance and pose of the individuals. In this case, the LLM generates the question: "What is the appearance of Person1 and Person4?" This question is designed to extract more contextually relevant details from the image captioner. The captioner then provides a refined description: "Person1 is wearing a suit with a gun and Person4 is wearing a suit with a gun." This additional information helps to better understand the scene and narrows down the possible answers to the original question. This detailed process highlights how our system leverages both multi-modal and textual information to generate precise and contextually relevant descriptions, ultimately improving the performance on complex visual reasoning tasks.

A.2 Qualitative CVR-ICL examples

Section 3.2 only illustrates the mechanism of our CVR-ICL. Here, we explain more details about its implementation. Figures 12 showcases one example of our CVR-ICL on the WinoGAViL (Bitton et al., 2022).



Question: Why do person1 and person4 have guns ?
Choices: 1) They are soldiers.
2) Person1 and person4 are robbing a hotel room.
3) They are cattle thieves.
4) They are about to shoot someone.

Please describe the image content in details ?

An image of a man in a suit with a gun and another in a suit with a gun.

{"role": "system", "content": "A chat between a curious human and an artificial intelligence assistant. You are a questioner for an image caption model and need to ask one question to get crucial information for answer prediction."}
{"role": "user", "content": "The captioner already generate a detailed image caption 'An image of a man in a suit with a gun and another in a suit with a gun'. Now you need to ask only one question for a special question 'why do person1 and person4 have guns' with chocies '1) They are soldiers. 2) Person1 and person4 are robbing a hotel. 3) They are cattle thieves. They are about to shoot someone'. What question you will ask? (only consider use what to set question)"}

What is the appearance of Person1 and Person4?

What is the appearance of person1 and person4

Person1 is wearing a suit with a gun and Person4 is wearing a suit with a gun.

Figure 11: The detailed illustration of our CaID process on VCR. Best viewed by zooming in.

To accurately calculate similarity scores using the cosine similarity function, we utilize BM25 (Robertson et al., 1995) for text encoding and BLIP2 multi-embedding (Li et al., 2023) for multi-modal inputs. As illustrated in Figure 12, the process begins with encoding both the test and training prompts through multi-modal and text-based encoders. For instance, a test case from WinoGAViL might contain the question "Select two pictures most related to clouds?" along with images of a foggy river, a cloud of sand on a beach, and other related scenes. At the beginning, the multi-modal encoder processes these images as well as the question and generates multimodal-level embeddings. Simultaneously, we convert these images into context-aware image descriptions and translate the entire case into text form. The text-based encoder then generates corresponding text-level embeddings. Next, we calculate the individual cosine similarity scores in both the multi-modal and text domains. The final similarity score, which determines the most relevant cases, is calculated in a balanced manner as $S = S_1 + S_2$. These scores are then sorted, and the top-k most similar cases are selected as in-context learning examples. This dual-encoding and similarity scoring approach ensures

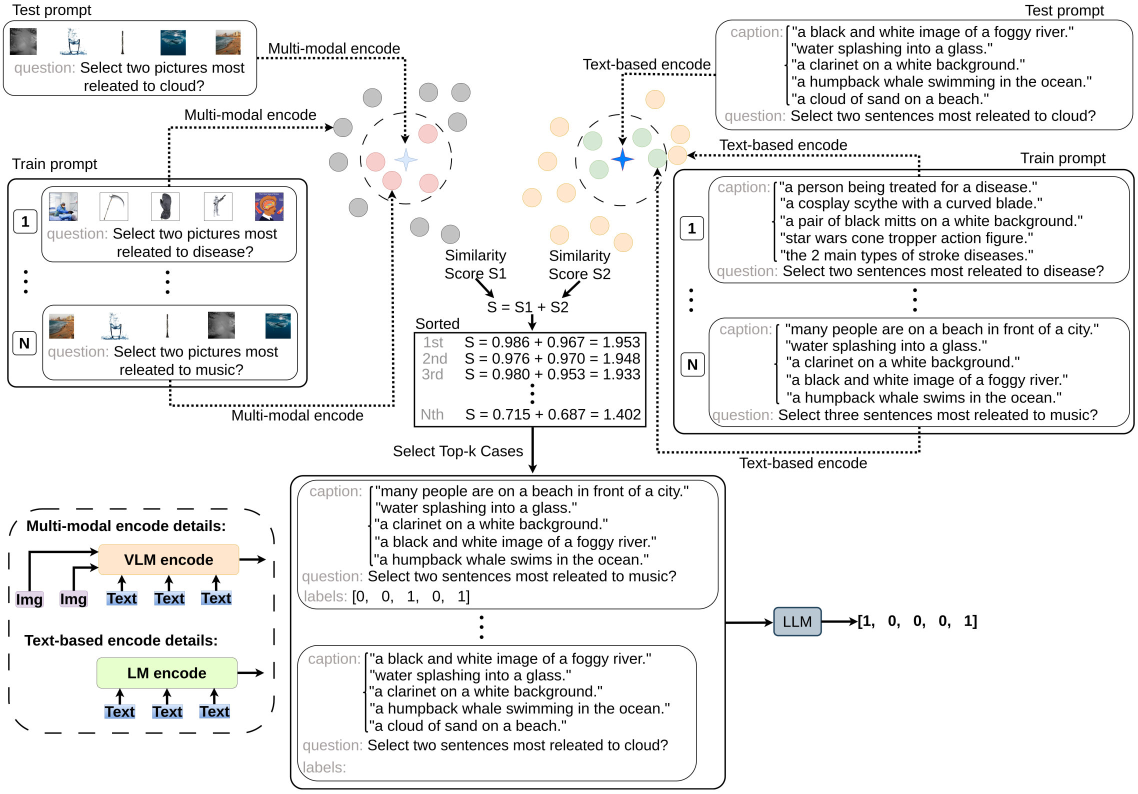

Test prompt
question: Select two pictures most releated to cloud?
Multi-modal encode
Multi-modal encode
Train prompt
1
question: Select two pictures most releated to disease?
N
question: Select two pictures most releated to music?
Multi-modal encode
Test prompt
caption: "a black and white image of a foggy river." "water splashing into a glass." "a clarinet on a white background." "a humpback whale swimming in the ocean." "a cloud of sand on a beach."
question: Select two sentences most releated to cloud?
Text-based encode
Text-based encode
Train prompt
1
caption: "a person being treated for a disease." "a cosplay scythe with a curved blade." "a pair of black mitts on a white background." "star wars cone tropper action figure." "the 2 main types of stroke diseases."
question: Select two sentences most releated to disease?
N
caption: "many people are on a beach in front of a city." "water splashing into a glass." "a clarinet on a white background." "a black and white image of a foggy river." "a humpback whale swims in the ocean."
question: Select three sentences most releated to music?
Text-based encode
Similarity Score S1
Similarity Score S2
S = S1 + S2
Sorted
1st S = 0.986 + 0.967 = 1.953
2nd S = 0.976 + 0.970 = 1.948
3rd S = 0.980 + 0.953 = 1.933
Nth S = 0.715 + 0.687 = 1.402
Select Top-k Cases
caption: "many people are on a beach in front of a city." "water splashing into a glass." "a clarinet on a white background." "a black and white image of a foggy river." "a humpback whale swims in the ocean."
question: Select two sentences most releated to music?
labels: [0, 0, 1, 0, 1]
caption: "a black and white image of a foggy river." "water splashing into a glass." "a clarinet on a white background." "a humpback whale swimming in the ocean." "a cloud of sand on a beach."
question: Select two sentences most releated to cloud?
labels:
LLM
[1, 0, 0, 0, 1]
Multi-modal encode details:
VLM encode
Img
Img
Text
Text
Text
Text-based encode details:
LM encode
Text
Text
Text

Figure 12: The detailed illustration of our CVR-ICL on WinoGAViL. Best viewed by zooming in.

that we capture the nuanced relationships between multi-modal inputs and text, thereby enhancing the accuracy and relevance of our in-context learning framework.

A.3 Comparative Analysis with Fine-tuned Models

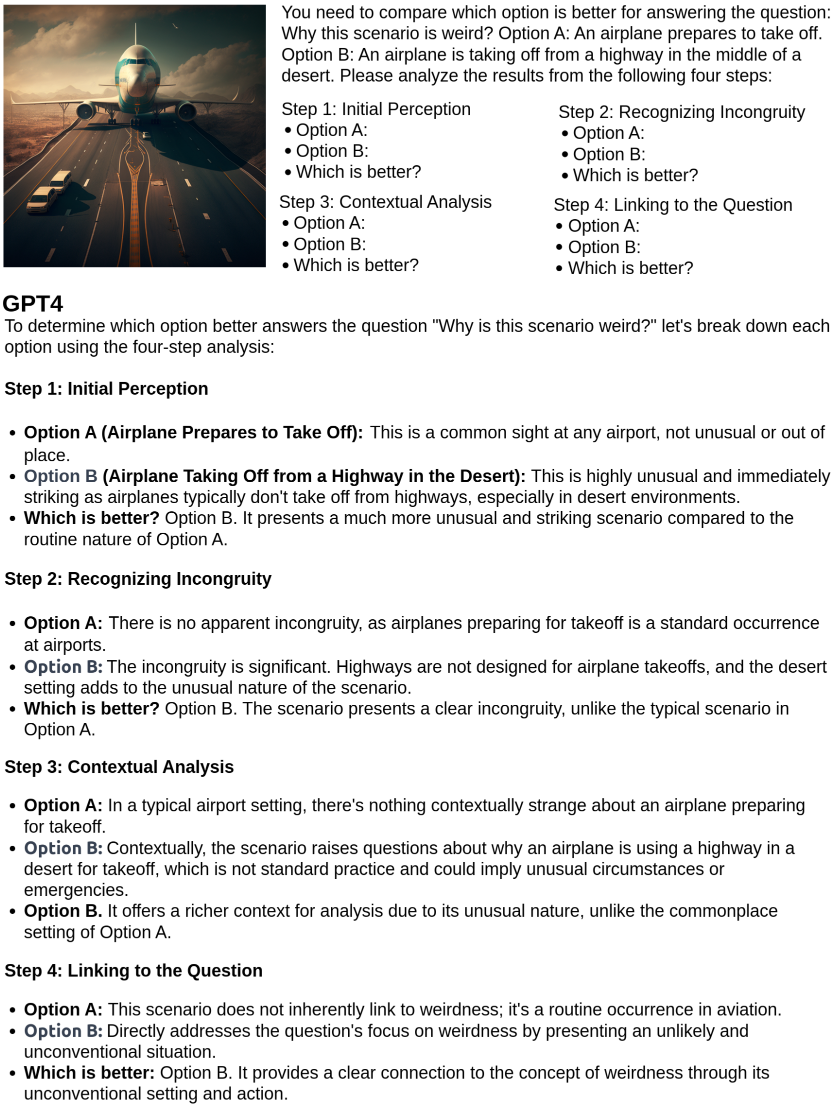
In this section, we explore the impact of fine-tuning strategy on performance in complex visual reasoning tasks. Since some tasks in the complex visual reasoning field are initially designed in the supervised setting, we are curious whether our approach can also perform better with the help of real annotation. For the test-only datasets WinoGAViL and Winoground, we randomly divided them into splits of 80% training, 10% validation, and 10% testing. Due to the small number of cases in these tasks, we abandoned training LLMs to avoid catastrophic forgetting. Instead, we choose to fine-tune the captioner using the real labels and incorporated these real annotations into our CVR-ICL examples. Results shown in Table 5 compare our CVR-LLM's performance in zero-shot and fine-tuned settings against SOTA performances, revealing that our method maintains SOTA performance in several areas.

Dataset	Category	Zero-shot		Finetuned	
		SOTA	CVR-LLM	SOTA	CVR-LLM
WinoGAViL	5/6	55.0	74.7	54.6	**82.8**
	10/12	52.0	73.2	47.2	**80.8**
	SWOW	59.0	88.7	68.8	**95.9**
Winoground	Text	46.5	43.5	47.0	**55.0**
	Image	27.7	35.0	42.2	**42.5**
	Group	24.2	26.5	30.5	**35.0**
Whoops	GPT-4 Rate	31.0	62.0	71.0	**72.0**
VCR	Q->A	48.8	52.9	**87.4**	85.3
	QA->R	49.7	54.3	**89.6**	87.5
	Q->AR	28.6	30.4	**78.6**	77.1
NYCCC	Match acc.	60.4	60.6	**84.5**	80.9
	CrowdAcc	59.2	57.4	**73.3**	69.6
	NYAcc	66.5	63.1	**68.2**	65.4

Table 5: The comparison of our CVR-LLM against SOTA performance under two kinds of settings.

A.4 More explanation about our CoC

The Chain-of-Comparison (CoC) is designed to qualitatively analyze the semantic contribution of context-aware image descriptions against general image captions. It is inspired by the popular idea of Chain-of-Thought, which implements a step-by-step analysis to evaluate effectiveness. Figure 13 shows an example from the Whoops dataset, comparing the semantic gap between a general caption "An airplane prepares to take off" (Option A) and our context-aware image description "An airplane is taking off from a highway in the middle of the



You need to compare which option is better for answering the question: Why this scenario is weird? Option A: An airplane prepares to take off. Option B: An airplane is taking off from a highway in the middle of a desert. Please analyze the results from the following four steps:

Step 1: Initial Perception
- Option A:
- Option B:
- Which is better?

Step 2: Recognizing Incongruity
- Option A:
- Option B:
- Which is better?

Step 3: Contextual Analysis
- Option A:
- Option B:
- Which is better?

Step 4: Linking to the Question
- Option A:
- Option B:
- Which is better?

GPT4

To determine which option better answers the question "Why is this scenario weird?" let's break down each option using the four-step analysis:

Step 1: Initial Perception

- **Option A (Airplane Prepares to Take Off):** This is a common sight at any airport, not unusual or out of place.
- **Option B (Airplane Taking Off from a Highway in the Desert):** This is highly unusual and immediately striking as airplanes typically don't take off from highways, especially in desert environments.
- **Which is better?** Option B. It presents a much more unusual and striking scenario compared to the routine nature of Option A.

Step 2: Recognizing Incongruity

- **Option A:** There is no apparent incongruity, as airplanes preparing for takeoff is a standard occurrence at airports.
- **Option B:** The incongruity is significant. Highways are not designed for airplane takeoffs, and the desert setting adds to the unusual nature of the scenario.
- **Which is better?** Option B. The scenario presents a clear incongruity, unlike the typical scenario in Option A.

Step 3: Contextual Analysis

- **Option A:** In a typical airport setting, there's nothing contextually strange about an airplane preparing for takeoff.
- **Option B:** Contextually, the scenario raises questions about why an airplane is using a highway in a desert for takeoff, which is not standard practice and could imply unusual circumstances or emergencies.
- **Option B.** It offers a richer context for analysis due to its unusual nature, unlike the commonplace setting of Option A.

Step 4: Linking to the Question

- **Option A:** This scenario does not inherently link to weirdness; it's a routine occurrence in aviation.
- **Option B:** Directly addresses the question's focus on weirdness by presenting an unlikely and unconventional situation.
- **Which is better:** Option B. It provides a clear connection to the concept of weirdness through its unconventional setting and action.

Figure 13: The detailed illustration of our CoC on Whoops. Best viewed by zooming in.

desert" (Option B).

Our CoC prompt asks the LLM to analyze the semantic contribution through four steps: Initial Perception, Recognizing Incongruity, Contextual Analysis, and Linking to the Question. This process mimics the human brain's analytical process. We directly ask the LLM to compare the contributions of the two options and determine which is better.

For instance, in the Initial Perception step, the LLM identifies Option B as superior because it is highly unusual and immediately striking, as airplanes typically do not take off from highways, especially in desert environments. This scenario is much more unusual and striking compared to the routine scenario of Option A, which merely depicts an airplane preparing to take off at an airport. During the Contextual Analysis step, Option B is again favored. The LLM explains that contextually, the scenario raises questions about why an airplane is using a highway in a desert for takeoff, which is not standard practice and could imply unusual circumstances or emergencies. Option A, in contrast, has nothing contextually strange about an airplane preparing for takeoff in a typical airport setting. Finally, in the Linking to the Question step, the LLM determines that Option B provides a clearer connection to the concept of weirdness through its unconventional and striking situation. Option A does not inherently link to weirdness, as it describes a routine occurrence in aviation.

This example demonstrates how our CoC framework effectively breaks down and evaluates the semantic contributions of different types of image descriptions, highlighting the advantages of context-aware image descriptions in complex visual reasoning tasks.