Boosting Visual Knowledge-Intensive Training for LVLMs Through Causality-Driven Visual Object Completion

Qingguo Hu[1,4] , **Ante Wang**[1] , **Jia Song**[1] , **Delai Qiu**[2] , **Qingsong Liu**[2] and **Jinsong Su**[1,3,4*]
[1]School of Informatics, Xiamen University, China
[2]Xiamen Unisound Intelligence Technology Co., Ltd
[3]Shanghai Artificial Intelligence Laboratory, China
[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
{huqingguo, wangante}@stu.xmu.edu.cn, jssu@xmu.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have experienced significant advancements in recent years. However, their performance still falls short in tasks requiring deep visual perception, such as identifying subtle differences between images. A potential cause is the scarcity of visual knowledge in popular instruction-tuning corpora, resulting in inadequate visual perception and reasoning capabilities. To address this challenge, we introduce a self-improvement framework grounded in a novel visual knowledge-intensive task, Causality-driven Visual object Completion (CVC). This task requires LVLMs to infer the masked object in an image based on its *causal* relationships with the other visible information. We first obtain rich examples cheaply through our automated instance construction pipeline, without relying on sophisticated LVLMs (*e.g.*, GPT-4V) or human assistance. Then, LVLMs effectively self-improve through trial and error learning using these created instances. Our experiments demonstrate substantial gains across four challenging specialized tasks and four widely-used comprehensive benchmarks. Especially on specialized tasks, our method achieves an average improvement of 5.4% and 4.0% compared to the corresponding baselines when utilizing LLaVA-1.5-7B and LLaVA-1.5-13B, respectively. Code and the supplementary file are available at https://github.com/XMUDeepLIT/CVC.

1 Introduction

In the pursuit of Artificial General Intelligence (AGI), the ability to effectively process and understand multimodal information is of paramount importance. The surge in Large Language Models (LLMs) [Achiam *et al.*, 2023] has catalyzed the development of powerful Large Vision-Language Models (LVLMs) [Yang *et al.*, 2023; Team *et al.*, 2023]. The main idea of most LVLMs is to integrate a pretrained visual encoder [Radford *et al.*, 2021] with LLMs using an alignment module [Liu *et al.*, 2023b]. By leveraging the inherent capabilities of LLMs, these LVLMs have achieved excellent performance in various visual tasks, such as image captioning [Yang *et al.*, 2023], visual question answering [Lan *et al.*, 2024a; Liu *et al.*, 2024a], and multimodal machine translation [Lin *et al.*, 2020; Lan *et al.*, 2023; Yin *et al.*, 2023; Lan *et al.*, 2024b].

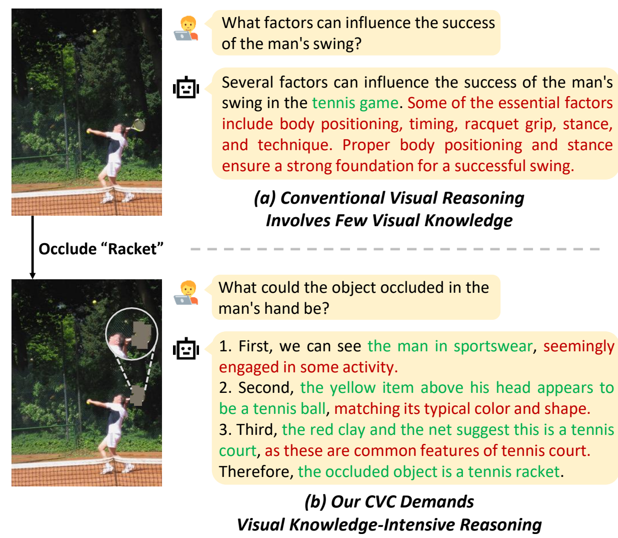

Figure 1: Comparison between conventional and our CVC instances for visual reasoning, where contents involving visual knowledge and linguistic commonsense knowledge are highlighted in green and red, respectively.

Despite the remarkable progress made by LVLMs, they still struggle with some basic visual perception and reasoning tasks, which humans can solve almost unerringly [Tong *et al.*, 2024; Fu *et al.*, 2024]. For instance, [Tong *et al.*, 2024] demonstrate that existing LVLMs may perceive images with clear visual differences as similar, thus failing to distinguish them. This issue may stem from the inherent deficiency of current LVLM training corpora, which prioritize commonsense knowledge over visual knowledge and lack complex reasoning tasks involving visual scenarios. In

*Corresponding Author

Figure 1(a), we showcase a typical example categorized as "*complex reasoning*" within the popular LLaVA training corpus [Liu *et al.*, 2023b]. The overall reasoning process involves limited visual knowledge ("*playing tennis*"), but instead, mostly relies on the LLM's intrinsic commonsense knowledge ("*the factors influencing the success of swing*"). This can be because widely-used corpora [Zhao *et al.*, 2023; Liu *et al.*, 2023b] are mostly crafted from strong language-only GPT-4 [Achiam *et al.*, 2023], thus naturally contain a large amount of linguistic knowledge. As a result, training with these instances only may limit the exploitation of LVLMs' capabilities in visual perception and reasoning.

To tackle this issue, we propose a novel self-improvement framework for further exploiting the visual capabilities of LVLMs autonomously. Specially, in addition to existing vision-language tasks [Goyal *et al.*, 2017; Krishna *et al.*, 2017; Kang *et al.*, 2023], we introduce causality-driven visual object completion (CVC), a challenging visual knowledge-intensive reasoning task for multimodal instruction tuning. This task is inspired by "*visual completion*" [Pessoa *et al.*, 1998] in perceptual psychology, where humans with high-level cognitive processing and reasoning skills are capable of extracting meaning even from incomplete visual information. Taking Figure 1(b) for example, given an image where an object is masked, the LVLM has to take the visible context as evidence and provide a step-by-step reasoning path (rationale) for explicitly inferring the masked object. Similar ideas have been successfully applied to the fields of vision-language for self-supervised pretraining [Chen *et al.*, 2020]. However, unlike these studies where the masked area is randomly selected, we especially emphasize the *high causality* between the masked object and its surroundings. This prevents the LVLM from forcibly fitting training targets that are uncertain and difficult to infer, encouraging it to conduct reasonable inference over the masked image.

To cost-effectively produce training instances of CVC, we first leverage widely available image-caption pairs [Lin *et al.*, 2014] and occlude the high-causality objects recognized in images. Particularly, the causality score between an object and its image context is empirically estimated by the confidence of a Masked Language Model (*e.g.*, RoBERTa [Liu *et al.*, 2019]) on its corresponding caption entity.

With the crafted CVC instances above, we aim to synthesize valid rationales that lead to the target answer for LVLM training. To address this challenge, we apply *trial and error learning* [Young, 2009] to LVLMs for self-improvement: for each CVC instance, we first ask an LVLM to synthesize multiple rationales (trials) for inferring the masked object. Then, the difficulty of each instance is assessed based on the frequency of its trials yielding the target answer. We select challenging instances that are valuable for training to enhance the LVLM's learning efficiency. Finally, these successful self-generated trials are fed to the LVLM for self-improvement. In this way, the LVLM is not only taught to recognize the detailed information of images (visual perception), but also encouraged to conduct "*slow thinking*" [Daniel, 2017] for explicitly predicting the masked object by leveraging its relevant surroundings (visual reasoning). Thus, the visual capabilities of the LVLM can be comprehensively improved without the help of humans or sophisticated LVLMs (*e.g.*, GPT-4V [Achiam *et al.*, 2023]).

To demonstrate the effectiveness of our proposed self-improvement framework, we conduct extensive experiments on challenging specialized tasks, including MMVP [Tong *et al.*, 2024], Winoground [Thrush *et al.*, 2022], V^*Bench [Wu and Xie, 2024], VSR [Liu *et al.*, 2023b] and comprehensive benchmarks: MME [Fu *et al.*, 2023], MMBench [Liu *et al.*, 2023c], SEEDBench [Li *et al.*, 2023] and MM-Vet [Yu *et al.*, 2023]. The results on LLaVA-1.5 family LVLMs show that our method consistently outperforms the corresponding baselines, particularly on the more challenging tasks of MMVP and Winoground, with improvements of +10.0% and +8.2%, respectively. Detailed analyses also indicate that our method can scale to larger data volumes, validating the promise of our method for more pervasive use.

2 Related Work

Large Vision-Language Models (LVLMs). Mainstream LVLMs adopt a similar architecture [Liu *et al.*, 2023b], where a vision encoder is linked to an LLM via an alignment module, enabling perception of visual information. Nowadays, LVLMs have shown excellent performance in downstream tasks. However, recent studies highlight their limitations in visual understanding. For instance, [Tong *et al.*, 2024] benchmark LVLMs on distinguishing fine-grained visual differences, revealing the poor performance of existing LVLMs. Similarly, [Fu *et al.*, 2024] point out that existing benchmarks overlook visual perception, and instead use classic vision tasks to assess LVLMs. Though easy for humans, these tasks remain challenging for current models.

Learning from Rationales. Early studies have demonstrated that human-annotated rationales can enhance model performance [Zhang *et al.*, 2023]. Today, thanks to the emerging reasoning abilities in LLMs [Wei *et al.*, 2022; Wang *et al.*, 2025], many studies [Ho *et al.*, 2022; Hsieh *et al.*, 2023] apply knowledge distillation to learn from synthetic rationales generated by advanced LLMs with hundreds of billions of parameters. However, the above approaches can be costly. Therefore, an alternative line of research focuses on self-improvement methodologies, enabling LLMs to learn from self-generated rationales [Huang *et al.*, 2022]. In the field of LVLMs, such self-improvement techniques remain under-explored. In this paper, we harness LVLMs' inherent reasoning abilities to autonomously generate rationales, using them to cheaply enhance the LVLM's visual capabilities.

Mask-then-Predict. This paradigm aims to enhance model performance by recovering masked signals. It has been extensively investigated in both language [Devlin *et al.*, 2018; Liu *et al.*, 2019] and vision [Bao *et al.*, 2021; He *et al.*, 2022] domains separately. Along this line, a series of works [Chen *et al.*, 2020; Kwon *et al.*, 2022] employ this paradigm to conduct Vision-Language Pre-training (VLP). Unlike conventional approaches in VLP where masked signals are directly predicted, our method requires the LVLM to infer the masked object through a step-by-step rationale. Additionally, we emphasize the high causality between the masked object and its

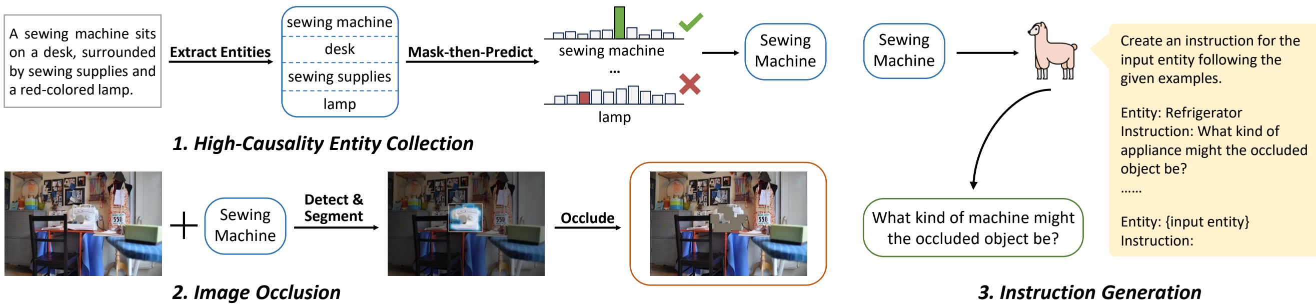

Figure 2: Overview of our data construction pipeline. Using widely available image-caption pairs, we construct CVC instances cost-effectively. Each CVC instance consists of a high-causality entity, an occluded image, and a task instruction.

surroundings, rather than applying random masking. By doing so, our method can further develop the LVLM's ability to utilize visible context and conduct reasonable inference over the masked image.

3 Our Framework

In this section, we introduce an innovative self-improvement framework for LVLMs. We start by providing a detailed formal definition of causality-driven visual object completion (CVC), a challenging visual knowledge-intensive reasoning task for LVLMs (§3.1). Next, we present an overview of the pipeline used to construct task instances cheaply (§3.2). Finally, using the constructed CVC instances, we utilize an LVLM to autonomously synthesize valid rationales. These self-generated data are used as additional training data for the multimodal instruction tuning of the LVLM (§3.3).

3.1 Task Definition

Formally, given an occluded image $I_{\setminus e}$, where an object e[1] having high causality with its surroundings is artificially masked, we ask an LVLM to infer the masked object by providing a step-by-step rationale r:

$$r = \texttt{LVLM}(I_{\setminus e}, q, p), \tag{1}$$

where the rationale r refers to natural language explanations that support the model's final prediction, q is the task instruction (*e.g.*, *"What is the occluded object?"*), and p is the Chain-of-Thought (CoT) prompt (*i.e.*, *"Let's think step by step"*) used to elicit the rationale. In this context, high causality refers to the strong logical association between the masked object and its surroundings, enabling reasonable inference.

Compared with predicting the object e only, providing a step-by-step rationale r can better help the LVLM learn to search for relevant cues in the visible context and then conduct causal reasoning over them to infer the answer. Therefore, deep visual perception and reasoning capabilities of the LVLM can be promoted through this process.

3.2 Data Preparation

Based on widely available image-caption pairs, we develop an automatic data construction pipeline to create the dataset of CVC, where each instance is composed of $\{e, I_{\setminus e}, q\}$. As shown in Figure 2, we successively construct the three elements of each instance through *(i) high-causality entity collection*, *(ii) image occlusion*, and *(iii) instruction generation*. We describe these steps in detail below.[2]

High-Causality Entity Collection. Directly estimating the causality between an object and its image context is non-trivial. Therefore, we empirically compute the causality score of an object by utilizing the uncertainty estimation of the corresponding entity in the caption, which shares the same semantic meaning as the corresponding image. Specifically, given an image-caption pair $\{I, T\}$, where I denotes the image and T represents the caption, we first employ LLaMA2-7B [Touvron *et al.*, 2023] to extract entities (*e.g.*, e_i) from T via in-context learning. Next, we mask an entity e_i in the caption T, yielding the masked text $T_{\setminus e_i}$. We then feed $T_{\setminus e_i}$ into RoBERTa [Liu *et al.*, 2019] (denoted as ϕ) to predict the masked tokens. The prediction probability of e_i serves as the causality score: $p(e_i \mid T_{\setminus e_i}; \phi)$. Intuitively, entities with high causality are more easily predicted based on their context, resulting in higher probabilities compared to low-causality entities. Finally, we select those high-causality entities with scores exceeding a threshold γ.

Image Occlusion. After collecting high-causality entities, we proceed to occlude these entities in their corresponding images. First, we utilize GLIP [Li *et al.*, 2022] to ground each entity's corresponding object in the image, thereby yielding a bounding box. This bounding box serves as input for SAM [Kirillov *et al.*, 2023], which predicts a mask that precisely delineates the region of the object. We then apply heavy occlusion to the segmented pixels. We use rectangular boxes instead of actual contours to occlude objects. This prevents the LVLM from relying on shapes for predictions thus hindering its effective use of surroundings for reasoning.

Instruction Generation. The task instruction aims to elicit the desired LVLM predictions. Given the diverse range of entities we collected, a fixed instruction (*e.g.*, "*What is the occluded object?*") might lead to ambiguous references, thereby reducing the data efficiency because it is hard for an LVLM to yield the answers exactly. Besides, it limits the data diversity,

[1] For the sake of clarity, we denote both an object in an image and its corresponding entity by e.

[2] Final processed data, along with more implementation details, will be provided in our supplementary file.

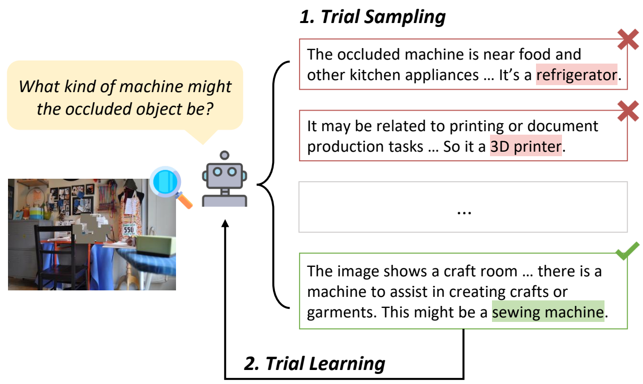

Figure 3: Overview of our self-improvement approach. Based on trial and error learning, the LVLM samples multiple rationales (trials) for a given CVC instance. The sampled trials are selectively learned based on their correctness and estimated difficulty.

which can be crucial for model generalization. Therefore, we construct a specialized instruction for each entity. To achieve this, we manually craft several examples of entity-instruction pairs and use LLaMA2-7B through in-context learning to generate a new instruction based on the input entity.

3.3 Model Training

Thus far, the crafted CVC instances still lacks valid rationales that lead to the target entity. To address this, we apply *trial and error learning* [Young, 2009], a core learning mechanism in behavioral science, to LVLMs for self-improvement: faced with an unfamiliar problem, the learner experiments with multiple trials, evaluates their successfulness based on environmental feedback, and learns from successful attempts to improve problem-solving capabilities. As shown in Figure 3, we illustrate how successful trials are sampled and then fed back into the LVLM for self-improvement.

Trial Sampling. Given each CVC instance $\{e, I_{\setminus e}, q\}$ and an LVLM parameterized by θ, each rationale produced by the LVLM for inferring the masked object is treated as a trial. We adopt the popular nucleus sampling [Holtzman *et al.*, 2019] to obtain N different trials:

$$r_1, r_2, \ldots, r_N \sim \texttt{LVLM}(I_{\setminus e}, q, p). \tag{2}$$

Then, we assess the successfulness of each trial r_j by verifying its corresponding answer $\hat{e}_j$ against the target entity e. Concretely, for each trial r_j, we extract the answer $\hat{e}_j$ utilizing LLaMA2-7B through in-context learning[3], and r_j is successful when $\hat{e}_j = e$.

Trial Learning. With the obtained successful trials for each instance, we only select challenging ones that are worth learning to train the LVLM for improving learning efficiency. To achieve this, we measure the difficulty of a CVC instance by calculating the inverse frequency of its trials that yield the target answer e: $\mathcal{F} = 1 - \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}(\hat{e}_k = e)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. Then, only successful trials from instances with difficulty scores higher than the threshold α are chosen. Finally, for an instance $\{e, I_{\setminus e}, q\}$ with a successful trial r', we train the LVLM by hybridizing the trial and target entity as supervised signals, using the following standard cross-entropy loss:

$$\mathcal{L} = -\left(\log p(e \mid I_{\setminus e}, q; \theta) + \log p(r' \mid I_{\setminus e}, q, p; \theta)\right). \tag{3}$$

The hybrid loss term promotes the "*fast and slow thinking*" of the LVLM via learning to not only predict the direct answer but also decompose the complex visual completion process step by step. To maintain the LVLM's instruction-following ability, we combine our data with general multimodal instruction data and jointly perform multimodal instruction tuning.

Through this process, without the need for sophisticated LVLMs (*e.g.*, GPT-4V) or human annotations, the LVLM autonomously learns from its own scarce but successful trials. While achieving proficiency in CVC, the LVLM's visual perception and reasoning capabilities are further developed.

4 Experiments

4.1 Setup

Evaluation Datasets. We conduct in-depth analyses on a range of challenging specialized tasks and widely-used comprehensive benchmarks, aiming to test the effectiveness of our method on the deep visual perception and general capabilities of LVLMs, respectively.

- **Challenging specialized tasks:** MMVP [Tong *et al.*, 2024], Winoground [Thrush *et al.*, 2022], V^*Bench [Wu and Xie, 2024], and VSR [Liu *et al.*, 2023a]. Among these, MMVP focuses on distinguishing fine-grained visual differences that LVLMs often overlook. Winoground tasks LVLMs to select the correct caption for an image from two options with easily confused spatial relationships. V^*Bench challenges the model to recognize subtle visual details, and VSR assesses visual spatial understanding capability.
- **Comprehensive benchmarks:** MME [Fu *et al.*, 2023], MMBench [Liu *et al.*, 2023c], SEEDBench [Li *et al.*, 2023], and MM-Vet [Yu *et al.*, 2023]. These benchmarks encompass a wide range of subtasks, providing a thorough assessment of our method's generalization.

We follow [Liu *et al.*, 2024a] to use the same testing scripts and evaluation metrics for fair comparison.

Implementation Details. Since most mainstream LVLMs share the same architecture, we follow [Zhou *et al.*, 2024] to choose popular LLaVA-1.5 [Liu *et al.*, 2024a] for experiments. Both the 7B and 13B versions are used for demonstrating the scalability of our method across various model sizes. The CVC instances are constructed based on COCO dataset [Lin *et al.*, 2014]. We set γ, N, and α to 0.3, 16, and 0.75, respectively. By default, we use 90K of our data for training across all experiments unless otherwise noted. During training, we combine our data with the 665K instruction data from LLaVA-1.5 for multimodal instruction tuning. To ensure a fair comparison, our training starts from the pre-trained (*i.e.*, not yet instruction-tuned) weights of LLaVA-1.5, following the same training hyperparameters. All experiments are conducted on 8 × A100 80G GPUs.

[3] We provide the prompting examples in our supplementary file.

Method	Specialized Task				Comprehensive Benchmark			
	MMVP	Winoground	V^*Bench	VSR	MME	MMBench	SEEDBench	MM-Vet
LLaVA [Liu *et al.*, 2023b]	6.0	-	35.6	-	809.6	38.7	37.0	25.5
InstructBLIP [Dai *et al.*, 2023]	16.7	-	34.0	-	1212.8	36.0	58.8	26.2
Gemini Pro [Team *et al.*, 2023]	40.7	-	48.2	-	1496.6	73.6	62.4	64.3
GPT-4V [Achiam *et al.*, 2023]	38.7	-	55.0	-	1409.4	81.0	69.1	67.7
Naive-VC	27.3	29.5	48.7	66.9	1473.8	65.6	66.3	30.8
LLaVA-1.5-7B [Liu *et al.*, 2024a]	20.7	25.3	47.6	66.9	1510.7	64.3	66.1	31.1
w/ CVC	**30.7**	**33.5**	**49.7**	**68.1**	**1519.5**	**66.6**	**67.1**	**34.1**
Δ	+10.0	+8.2	+2.1	+1.2	+8.8	+2.3	+1.0	+3.0
LLaVA-1.5-13B [Liu *et al.*, 2024a]	33.3	34.3	48.7	67.3	1531.3	67.7	68.2	36.1
w/ CVC	**36.0**	**39.0**	**51.8**	**72.7**	**1543.8**	**70.1**	68.2	**38.2**
Δ	+2.7	+4.7	+3.1	+5.4	+12.5	+2.4	+0.0	+2.1

Table 1: Performance of our method on LLaVA-1.5 across all evaluation datasets. Baseline results are primarily sourced from [Tong *et al.*, 2024], [Wu and Xie, 2024] and [Liu *et al.*, 2024a]. Results for GPT-4V and Gemini Pro on comprehensive benchmarks are obtained from the respective official leaderboards.[4]

Method	MMBench			SEEDBench		
	LR	RR	AR	VR	AP	II
LLaVA-1.5-13B	44.1	62.6	70.4	77.0	38.6	73.2
LLaVA-1.5-7B	30.5	53.0	**73.4**	76.7	33.7	69.1
w/ CVC	**33.1**	**64.4**	70.4	**77.6**	**37.0**	**70.1**
Δ	+2.6	+11.4	-3.0	+0.9	+3.3	+1.0

Table 2: Results on visual reasoning tasks. The abbreviations for these tasks are as follows: Logical Reasoning (LR), Relation Reasoning (RR), and Attribute Reasoning (AR) for MMBench; and Visual Reasoning (VR), Action Prediction (AP), and Instance Interaction (II) for SEEDBench.

Baselines. To better validate the effect of our framework, we introduce a baseline termed Naive-VC. It is based on LLaVA-1.5-7B and trained extensively on data of the naive visual completion task, where entities are *randomly* selected before masking the corresponding objects in images. Additionally, we include results from other open-source LVLMs like LLaVA [Liu *et al.*, 2023b] and InstructBLIP [Dai *et al.*, 2023], as well as state-of-the-art models like GPT-4V [Achiam *et al.*, 2023] and Gemini Pro [Team *et al.*, 2023], to further demonstrate the effectiveness of our method.

4.2 Main Results

Our Method Notably Boosts Deep Visual Perception. As shown in Table 1, our method achieves notable improvements on 4 challenging specialized tasks. For LLaVA-1.5-7B, our method achieves nearly a 10% improvement on the two difficult tasks of Winoground and MMVP. Consistent performance gains are also significant in V^*Bench and VSR. Compared with Naive-VC, our method achieves more pronounced improvements. It is because CVC encourages the LVLM to engage in reasonable inference by leveraging richer contextual cues, demonstrating the importance of leveraging high-causality objects for visual completion. Notably, LLaVA-1.5-7B equipped with CVC data achieves performance comparable to or even surpassing vanilla LLaVA-1.5-13B on these tasks. For LLaVA-1.5-13B, our method also achieves substantial performance gains, with improvements of nearly 2%∼5%. These results demonstrate that our method can effectively enhance this more advanced LVLM, showcasing its potential to improve well-developed models at larger scales. Regarding comprehensive benchmarks, although our method is task-agnostic, it effectively enhances the performance of both LLaVA-1.5-7B and LLaVA-1.5-13B, achieving average improvements of 1.69% and 1.28%, respectively. Given that these tasks do not emphasize deep perception but involve a broader range of visual knowledge, these results indicate that our method effectively injects intensive visual knowledge into LVLMs and further aligns it within the language domain, demonstrating the generalizability of our approach.

[4]We provide the URL for the leaderboard of each comprehensive benchmark in our supplementary file.

Our Method Further Enhances Reasoning in Visual Scenarios. Since our method also encourages the LVLM to engage in visual reasoning, we further investigate how it impacts the reasoning ability of the LVLM in visual scenarios. To this end, we assess performance on several reasoning subtasks from MMBench and SEEDBench, which require not only basic perception but also cognitive inference [Li *et al.*, 2023; Liu *et al.*, 2023c]. As shown in Table 2, the performance of LLaVA-1.5-7B is generally improved on these reasoning tasks. Notably, the performance on relation reasoning exhibits the most substantial improvement, with an increase of 11.4%. These results indicate that our method effectively enhances the LVLM's reasoning ability, further validating its effectiveness from another perspective.

4.3 Analysis

Effect of Scaling Up CVC Data. We further investigate the effect of CVC data scale on downstream visual tasks. We apply our method to LLaVA-1.5-7B with varying CVC data scales, ranging from 30K to 150K. To highlight the effectiveness of our method on the deep visual perception, we focus on three challenging tasks: MMVP, Winoground and V^*Bench. The results are shown in Figure 4. We can observe

Synthesizer	Recall (%)	Specialized Task				Comprehensive Benchmark			
		MMVP	Winoground	V^*Bench	VSR	MME	MMBench	SEEDBench	MM-Vet
MiniGPT-v2	**17.3**	**32.7**	<u>32.3</u>	**52.9**	<u>68.7</u>	<u>1516.5</u>	65.8	<u>67.4</u>	32.7
LLaVA-1.5-7B	<u>12.1</u>	<u>30.7</u>	**33.5**	<u>49.7</u>	68.1	**1519.5**	**66.6**	67.1	**34.1**
LLaVA-NeXT-34B	10.3	<u>30.7</u>	31.8	48.7	**70.8**	1497.8	<u>66.3</u>	**67.6**	<u>33.2</u>

Table 3: Comparison of different LVLMs used as the rationale synthesizer. All training is conducted on LLaVA-1.5-7B. Recall indicates the proportion of successful rationales generated by each synthesizer. The best results are bolded, and the second-best results are underlined.

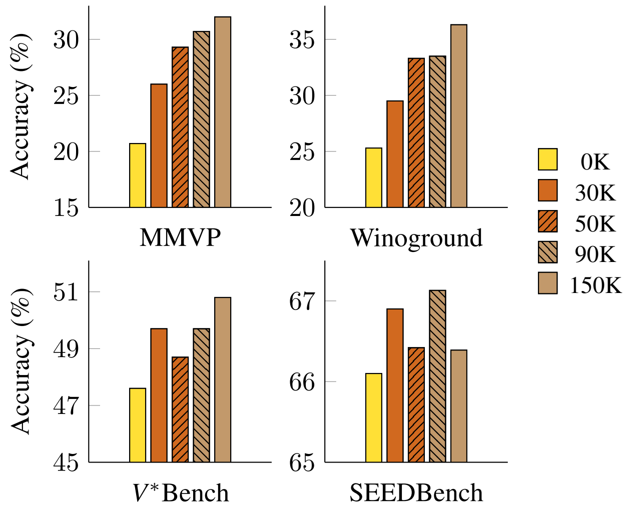

Figure 4: Results of scaling up CVC data for LLaVA-1.5-7B.

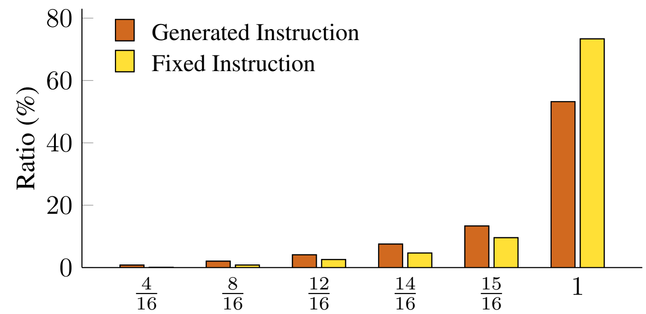

Figure 5: Comparison between two different task instruction settings in the distribution of $\mathcal{F}$ across all CVC instances. Both settings use LLaVA-1.5-7B for rationale synthesis.

a consistent trend of performance improvement across three tasks as the CVC data scale increases. Notably, after training with 150K of CVC data, our method achieves an accuracy of 36.3% on Winoground and 32.0% on MMVP, showing absolute improvements of 11.0% and 11.3% compared to the vanilla LLaVA-1.5-7B. A positive impact of CVC data scale on V^*Bench is also observed, though it is relatively less pronounced. Analyzing the effect of each data scale setting, we find that the 30K setting yields the most significant incremental improvement across the three tasks. After exceeding 90K, the rate of performance improvement slows, but the LVLM continues to benefit from additional data. However, in general scenarios (e.g., SEEDBench), we observe a slight performance decline when scaling up to 150K. This may be attributed to the excessive proportion of CVC data in the training corpus, which diminishes the impact of general data. Nonetheless, from the perspective of core visual perception, these results demonstrate that our method is not only efficient with limited data but also scalable to larger data volumes.

Influence of Different Synthesizers. So far, we have shown the effectiveness of our method in self-improvement. Nevertheless, other LVLMs can also function as the rationale synthesizer within our framework. In Table 3, we present the performance of MiniGPT-v2 [Chen *et al.*, 2023] and LLaVA-NeXT-34B [Liu *et al.*, 2024b] as the synthesizer. Specifically, MiniGPT-v2 is an LVLM based on LLaMA2-7B, and LLaVA-NeXT-34B is a more advanced LVLM trained on superior multimodal data. The results first reveal that CVC presents a substantial challenge to current LVLMs, as evidenced by the extremely low recall of each synthesizer. Second, our results show that different synthesizers yield comparable improvements. We attribute this to our method ensuring comparable data quality across the synthesizers by validating rationales with gold answers. As a result, our method is still effective when applied to less sophisticated LVLMs. Third, the improvement of our method in deep visual perception is closely tied to the synthesizer's performance on CVC. For example, MiniGPT-v2, with the highest recall of 17.3%, exhibits overall more advanced performance on 4 challenging specialized tasks. In contrast, although LLaVA-NeXT-34B has achieved state-of-the-art results across a wide range of benchmarks [Liu *et al.*, 2024b], its lower proficiency in CVC results in relatively modest improvements. This suggests that utilizing LVLMs with higher proficiency in CVC could further enhance the potential of our method. We hope these findings offer valuable insights for future research.

Generated Instruction Promotes Data Diversity. To evaluate the effect of generated instruction, we compare this setting with using a fixed instruction ("*What is the occluded object?*"). First, our results demonstrate that using generated instructions significantly outperforms a fixed instruction in terms of data diversity. As shown in Figure 5, using a fixed instruction results in nearly 20% more CVC instances with no successful rationale compared to using generated instructions. This indicates that generated instruction can lead to more specific references, thereby aiding the LVLM in producing accurate answers. Due to this design, our data covers nearly 2,000 different entities, ensuring data diversity. Second, in Table 4, using generated instructions exhibits better performance on both specialized and general visual tasks.

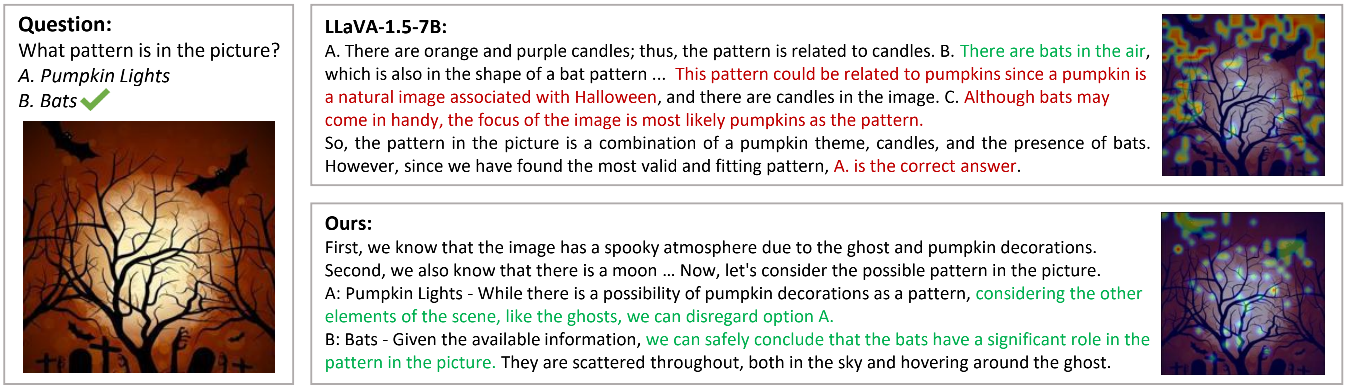

Figure 6: An example comparison of LLaVA-1.5-7B with and without our method. **Left**: A test sample from MMVP includes a question and an image, with the correct option annotated. **Right**: The reasoning processes and relative attention maps of the two methods. The highlighted regions in the attention map reveal where this method places more attention compared to the other.

Method	MMVP	Wino ground	SEED Bench	MM-Vet
LLaVA-1.5-7B	20.7	25.3	66.1	31.1
w/ CVC ($\alpha = 0.5$)	30.0	29.5	66.9	31.7
w/ CVC ($\alpha = 0.75$)	**30.7**	**33.5**	**67.1**	**34.1**
w/o Generated Inst.	28.7	32.5	67.1	31.8

Table 4: Results of our method enhanced by introducing generated instructions (Generated Inst.) and harder CVC instances. We conduct this ablation study on LLaVA-1.5-7B.

Rationale	Direct Answer	MMVP	Wino ground	SEED Bench	MM-Vet
✗	✗	20.7	25.3	66.1	31.1
✓	✗	28.0	26.0	66.8	29.4
✗	✓	30.7	27.0	66.3	29.2
✓	✓	**30.7**	**33.5**	**67.1**	**34.1**

Table 5: Ablation study of the hybrid loss on LLaVA-1.5-7B.

Harder CVC Instances Boost Greater Performance. We measure the difficulty of a CVC instance based on the frequency of its trials that yield the target answer. To investigate how the difficulty of CVC instances impacts LVLM performance, we collect CVC instances with different difficulties for LVLMs training. As depicted in Table 4, the harder CVC instances ($\alpha = 0.75$) exhibit a more pronounced performance improvement compared to the easier ones ($\alpha = 0.5$). This finding underscores that training with harder CVC instances drives the LVLM to achieve more substantial gains in tackling complex visual scenarios, highlighting the importance of instance difficulty in boosting model performance.

Effect of Hybrid Loss. Another key decision of our method is how to format the synthesized data for LVLMs training. To investigate this, we conduct an ablation study of the hybrid loss. As shown in Table 5, both rationale-only and direct-answer-only training can generally enhance the LVLM's performance on downstream visual tasks, with the exception of MM-Vet. When combining both rationale and direct answer for training, the hybrid loss enables the LVLM to engage in "*fast and slow thinking*," leading to further improvements in overall performance.

4.4 Case Study

We present a case study to underscore the efficacy of our method in visual reasoning that demands deep perception. As shown in Figure 6, although vanilla LLaVA-1.5-7B successfully identifies the key visual features (*"bats"*), it tends to rely on its linguistic commonsense knowledge (*i.e.*, *"pumpkin lights"* is a more Halloween-related pattern), leading to an incorrect inference. In contrast, our method enables the LVLM to confidently harness its visual perception capability, accurately identify the key pattern (*"bats"*) and conduct a reasonable inference over the image. Additionally, by comparing the relative attention maps of the two methods, we observe that our method places most attention on the key visual features, demonstrating a more precise and focused attention pattern. Conversely, the vanilla LLaVA-1.5-7B exhibits a more dispersed attention distribution, failing to effectively isolate the key visual features. These results further validate that our method effectively promotes the visual perception and reasoning capabilities of LVLMs.

5 Conclusion

In this paper, we propose a self-improvement framework that autonomously enhances the visual perception and reasoning capabilities of LVLMs. This framework is grounded in causality-driven visual object completion (CVC), which requires LVLMs to perform visual knowledge-intensive reasoning. We first develop a pipeline for constructing high-causality CVC instances. Then, leveraging trial-and-error learning, we harness the LVLM's inherent reasoning ability to synthesize rationales for each CVC instance and select more challenging ones for self-improvement. Experiments conducted on both challenging specialized tasks and comprehensive benchmarks demonstrate that our method significantly enhances the visual capabilities of LVLMs, particularly in scenarios demanding deep visual perception and reasoning.

Acknowledgments

The project was supported by National Key R&D Program of China (No.2022ZD0160501), Natural Science Foundation of Fujian Province of China (No.2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

Contribution Statement

Qingguo Hu and Ante Wang contributed equally to this work.

References

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

[Bao *et al.*, 2021] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

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

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

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

[Daniel, 2017] Kahneman Daniel. *Thinking, fast and slow*. 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[Fu *et al.*, 2024] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

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

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. of CVPR*, 2022.

[Ho *et al.*, 2022] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

[Holtzman *et al.*, 2019] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[Hsieh *et al.*, 2023] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[Huang *et al.*, 2022] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[Kang *et al.*, 2023] Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. Bigvideo: A large-scale video subtitle translation dataset for multimodal machine translation. *arXiv preprint arXiv:2305.18326*, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. of CVPR*, 2023.

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

[Kwon *et al.*, 2022] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022.

[Lan *et al.*, 2023] Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. Exploring better text image translation with multimodal codebook. *arXiv preprint arXiv:2305.17415*, 2023.

[Lan *et al.*, 2024a] Zhibin Lan, Liqiang Niu, Fandong Meng, Wenbo Li, Jie Zhou, and Jinsong Su. Avg-llava: A large multimodal model with adaptive visual granularity. *arXiv preprint arXiv:2410.02745*, 2024.

[Lan *et al.*, 2024b] Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. Translatotron-v(ison): An end-to-end model for in-image machine translation. *arXiv preprint arXiv:2407.02894*, 2024.

[Li *et al.*, 2022] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proc. of CVPR*, 2022.

[Li *et al.*, 2023] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of ECCV*, 2014.

[Lin *et al.*, 2020] Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In *Proc. of ACMMM*, 2020.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2023a] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2023.

[Liu *et al.*, 2023c] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proc. of CVPR*, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[Pessoa *et al.*, 1998] Luiz Pessoa, Evan Thompson, and Alva Noë. Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and brain sciences*, 1998.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 2021.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Thrush *et al.*, 2022] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proc. of CVPR*, 2022.

[Tong *et al.*, 2024] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proc. of CVPR*, 2024.

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

[Wang *et al.*, 2025] Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. Litesearch: Efficient tree search with dynamic exploration budget for math reasoning. In *Proc. of AAAI*, 2025.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 2022.

[Wu and Xie, 2024] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proc. of CVPR*, 2024.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[Yin *et al.*, 2023] Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. Multi-modal graph contrastive encoding for neural machine translation. *Artificial Intelligence*, 2023.

[Young, 2009] H Peyton Young. Learning by trial and error. *Games and economic behavior*, 2009.

[Yu *et al.*, 2023] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[Zhang *et al.*, 2023] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[Zhao *et al.*, 2023] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.

[Zhou *et al.*, 2024] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024.