

VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models

Zejun Li^{1†}, Ruipu Luo^{2†}, Jiwen Zhang¹, Minghui Qiu²,
Xuanjing Huang³, Zhongyu Wei^{1,4,5*}

¹School of Data Science, Fudan University

²ByteDance

³School of Computer Science, Fudan University

⁴Research Institute of Intelligent and Complex Systems, Fudan University

⁵Shanghai Innovation Institute

{zejunli20, zywei}@fudan.edu.cn

Abstract

While large multi-modal models (LMMs) have exhibited impressive capabilities across diverse tasks, their effectiveness in handling complex tasks has been limited by the prevailing single-step reasoning paradigm. To this end, this paper proposes **VoCoT**, a multi-step Visually-grounded object-centric Chain-of-Thought reasoning framework tailored for inference with LMMs. VoCoT is characterized by two key features: (1) object-centric reasoning paths that revolve around cross-modal shared object-level information, and (2) visually grounded representation of object concepts in a multi-modal interleaved and aligned manner, which effectively bridges the modality gap within LMMs during long-term generation. To adapt LMMs in reasoning with VoCoT, we further construct an instruction-tuning dataset. By combining VoCoT with the prevalent open-source LMM architectures, we develop a VoCoT-based model, **VolCano**. With only 7B parameters and limited input image resolution, VolCano demonstrates excellent performance across various scenarios. In benchmarks like CLEVR and EmbSpatial, which highly require complex reasoning capabilities, VolCano outperforms SOTA models, including powerful GPT-4V. Related code, models, and datasets are released in <https://github.com/RupertLuo/VoCoT>.

1 Introduction

In recent years, the success of large language models (LLMs) (OpenAI, 2023a,b) has been gradually extended to the multi-modal domain. By equipping LLM backbones (Touvron et al., 2023a,b; Chiang et al., 2023) with visual encoders (Radford et al., 2021) and efficient cross-modal alignment through generative training on image-text data (Liu et al., 2024b; Schuhmann et al., 2021), the constructed large multi-modal models (LMMs) possess the ca-

pabilities to perceive visual signals and engage in dialogue with users in multi-modal contexts (Liu et al., 2024b; Dai et al., 2023; Bai et al., 2023).

Despite the potential demonstrated by LMMs to serve as unified and versatile foundations, even models like GPT-4V struggle in composite tasks requiring complex analysis (Yang et al., 2023b; Wu and Xie, 2023), such as spatial reasoning (Du et al., 2024). We attribute this phenomenon to a major limitation of current LMMs: the prevailing single-step question-to-answer (Q2A) inference paradigm that directly generates answers based on questions (Dai et al., 2023; Liu et al., 2023b). As illustrated in Figure 1, correctly answering the question relies on analyzing the actions and relationships of multiple objects and thinking step-by-step, which is almost impossible to accomplish in a single-step prediction. Moreover, the single-step Q2A paradigm obscures the problem-solving process, limiting the interpretability of the LMM outputs. Conversely, in the language domain, the chain-of-thought (CoT) paradigm, which involves multi-step reasoning, has been widely explored in LLMs (Kojima et al., 2022; Wei et al., 2022), indicating a promising way for enhancing LMMs.

However, for complex contexts where multi-modal information coexists, constructing effective multi-step reasoning paths faces several challenges: (1) **Difficulty in integrating reasoning anchors within multi-modal contexts**. Textual CoTs mainly extract key information from contexts, such as entities, as anchors and conduct multi-step reasoning around these anchors (Yao et al., 2024). In multi-modal contexts, the anchor information is further required to be concepts shared between images and texts and establish connections between modalities. Existing works either supplement the image with additional information (such as segmentation maps (Yang et al., 2023a) and dot grids (Lei et al., 2024)) as anchors, but such information can only be

*Corresponding authors.

†Equal contribution.

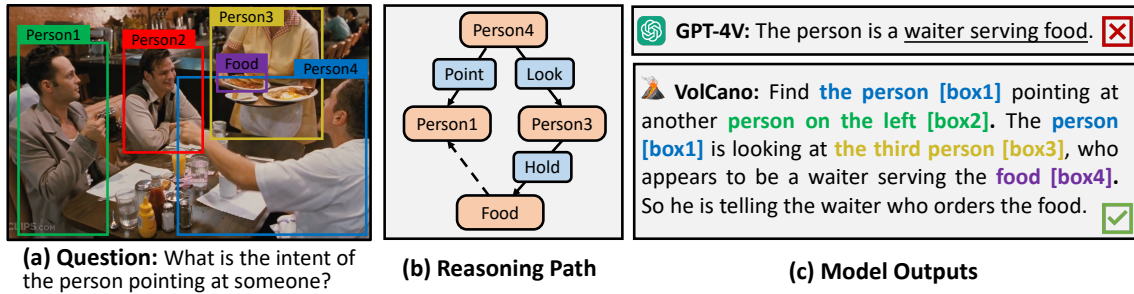


Figure 1: An example to compare different inference paradigms in LMMs. (a) A visual question that requires complex reasoning. (b) The conceptual object-centric reasoning path constructed to solve the problem. (c) Outputs of GPT-4V and the proposed VolCano. Hallucination is included in the output of GPT-4V. VolCano performs multi-step reasoning in the VoCoT format. In the reasoning path, key objects are highlighted and colors indicate the correspondence between object descriptions and the grounded regions in the image. “[box]” represents the coordinates of mentioned objects. Visual representations of objects are omitted for brevity.

effectively utilized by GPT-4V, or they roughly consider a single region as the anchor through a search-based approach (Shao et al., 2024), failing to model complex multi-object interactions. (2) **Limited grounding capabilities of LMMs.** During the generation process, LMMs may fail to ground textual descriptions to the corresponding visual information, resulting in erroneous information generated. For example, GPT-4V incorrectly ground the target person to the waiter in Figure 1. The risk of hallucination (Li et al., 2023d; Wang et al., 2023b) further hinders effective multi-step reasoning.

To address these challenges, we introduce a framework to empower LMMs for effective and reliable multi-step reasoning. We propose **VoCoT**, Visually grounded object-centric Chain of Thought. VoCoT is a CoT format that is compatible with LMM inference: (1) As illustrated in Figure 1 (a, b), objects serve as fundamental semantic units in both images and text, effectively bridging multi-modal information. Therefore, **VoCoT leverages objects as anchors for reasoning.** LMMs are encouraged to conduct multi-step analysis on the properties of key objects, as well as the relationships between them, ultimately reaching a conclusion. (2) To ensure the reliability of reasoning paths, **VoCoT represents objects in a visually grounded format:** a tuple of <textual description, coordinates, corresponding visual representations>. Models are required to explicitly ground objects in images by generating coordinates for them. Visual representations of objects are supplemented to enhance the cross-modal relevance in reasoning paths. This design mimics the habit of human, where we continuously reference the visual information of an object in the image when we mention it. (3) We

propose a **RefBind mechanism** to efficiently obtain the representations of objects without extra computation. Specifically, RefBind indexes the representation of each object from the image representation based on its coordinates. Generally, VoCoT constructs multi-modal interleaved reasoning paths where cross-modal aligned anchors are incorporated as shown in Figure 1 (c).

Nevertheless, there is a significant disparity between VoCoT and the formats of existing visual instruction data. To this end, we further construct a dataset, VoCoT-Instruct-80K, to train LMMs for reasoning in the format of VoCoT. VoCoT-Instruct-80K is built on multiple data sources: (1) Verbalizing structured reasoning paths from GQA (Hudson and Manning, 2019). (2) Supplementing visual QA pairs with thought processes. (3) Constructing complex questions and reasoning paths from images annotated with objects. By curating a wide range of data and leveraging assistance of GPT-4V, the presented dataset maintains both diversity and consistency in the desired format.

Based on the introduced VoCoT framework and dataset, we develop **VolCano**, a Visually-grounded multi-modal Chain-of-thought reasoning model. With only 7B parameters and 336^2 input resolution, VolCano excels in various scenarios and even surpasses GPT-4V on benchmarks like CLEVR (Johnson et al., 2017) and EmbSpatial (Du et al., 2024) that highly require complex reasoning.

2 Visually-grounded Object-centric CoT

In this section, we explain how to enable LMM to perform multi-step reasoning in the format of visually-grounded object-centric chain-of-thought

(VoCoT). In Section 2.1, we elaborate on the formulation of VoCoT. In Section 2.2, we present how to transform existing data resources into instruction-tuning datasets aligned with the VoCoT format.

2.1 VoCoT Formulation

VoCoT requires LMMs to perform step-by-step reasoning based on the provided context. Following textual CoTs (Kojima et al., 2022; Wei et al., 2022; Wang et al., 2022), the reasoning logic in VoCoT is primarily expressed in text but is not limited to specific formats. However, there exists a significant gap between multi-modal and text-only contexts. In order to construct effective and reliable reasoning paths in multi-modal contexts, we characterize VoCoT with two features: (1) **Object-centric**. Objects are the basic semantic units in images and can serve as anchors to establish connections between multi-modal contextual information. Therefore, VoCoTs are required to include important objects, followed by relevant information extraction and analysis. (2) **Visually-grounded**. Key objects included in VoCoT should be represented by tuples of $\langle \text{text description}, \text{coordinates}, \text{visual object representation} \rangle$. During inference, LMMs are required to generate both text and coordinates for objects to explicitly ground them within the images. The visual representation of objects further enhances the cross-modal relevance in the reasoning paths. Section 3.1 introduces how to obtain the visual object representations within current LMM frameworks.

2.2 VoCoT-Instruct-80K Dataset

The community has witnessed a surge in multi-modal instruction-following datasets (Liu et al., 2024b; Luo et al., 2023; Chen et al., 2024). However, none of these datasets meet the requirements of the VoCoT format, which includes responses to instructions (1) with CoT-formatted multi-step reasoning processes and (2) with visually grounded object-centric information, i.e., objects with corresponding coordinates. In this section, we introduce the pipeline to construct a VoCoT-formatted dataset from three types of existing data sources.

Type 1: GQA Source GQA (Hudson and Manning, 2019) is a VQA dataset that includes structured information: each image is paired with a scene graph, and a SQL-like reasoning path over the scene graph is provided for each VQA pair. An example is shown by the first part of Table 6 in Appendix A.1. Inspired by Shikra (Chen et al.,

2023b), we use a rule-based method to verbalize the SQL-like statements “[SematicStr]” and answers “[FullAnswer]” into fluent textual thoughts, supplementing objects descriptions with the corresponding coordinates from the scene graph.

Type 2: VQA-Based Source Another intuitive way to construct data in VoCoT format is to supplement VQA data with multi-step reasoning processes in the middle of the Q2A process. With the assistance of GPT-4V, reasoning thoughts are generated based on images, questions, answers, and object information within the images. The second part of Table 6 provides an example. Furthermore, we control the output format through in-context learning. Specifically, a crafted sample is included in the input context. As overly simple questions may not require complex reasoning, we sample a subset of data from complex reasoning problems in LLaVA-Instruct (Liu et al., 2024a) as the source.

Type 3: Image-Only Source Although the aforementioned two construction methods are effective, the generated data is limited to existing questions. To enhance the richness of questions and reasoning logic, we leverage GPT-4V to expand the constructed dataset. As illustrated by Table 6 in Appendix A.1, GPT-4V is provided with images and object information and prompted to generate complex questions, along with VoCoT-formatted reasoning paths and answers. In-context samples are also incorporated to ensure the correct output format. We choose LVIS (Gupta et al., 2019) as the data source due to the diversity of objects included.

Ultimately, we construct VoCoT-Instruct-80K, comprising 72K, 6K, and 2K samples from data sources of Type 1, 2, and 3, respectively. More details about the construction process, including the rule-based conversion approach, prompts for GPT-4V, in-context samples, and quality control methods used are provided in Appendix A.1.

3 VolCano: A VoCoT-enhanced LMM

In this section, we introduce how to adapt a modern LMM to utilize the VoCoT framework. We present the architecture of VolCano in Section 3.1, and detail the model training process in Section 3.2.

3.1 Architecture

As presented in Figure 2, the overall architecture of VolCano mainly follows LLaVA (Liu et al., 2023b). VolCano is built on top of a decoder-only LLM as

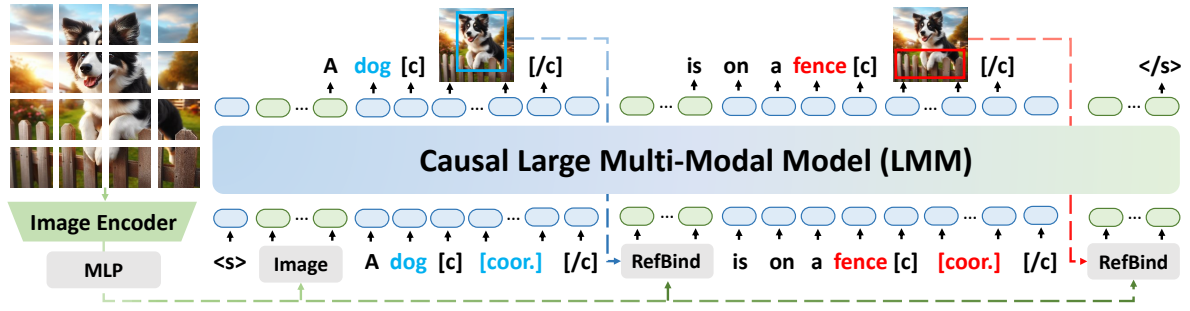


Figure 2: Illustration of the VolCano framework. The input and output are shown below and above the model, respectively. The blue and green rounded rectangles represent textual and visual tokens, respectively. Special tokens “[c]” and “[/c]” denotes the beginning and end of the coordinates (“[coor.]” in the figure). Coordinates are represented in text. In the output, we visualize coordinates by drawing corresponding boxes in the image for a better illustration. RefBind obtains the representations of objects with the image features and predicted coordinates.

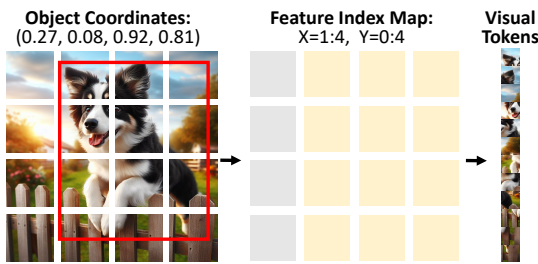


Figure 3: Illustration of the RefBind mechanism.

the backbone. We incorporate a vision transformer (ViT) as the visual encoder to encode image inputs. A two-layer MLP is adopted as the connection module to map the output of the visual encoder into the input space of the language backbone.

Representations of Multi-modal Sequences

VolCano represents image-text data as an interleaved sequence of visual and textual tokens. Text inputs are tokenized and represented using the embedding layer. Images and objects can appear at any position in the sequence and are represented by visual tokens. Images are encoded by the ViT. The obtained 2D feature maps are flattened into 1D sequences and further mapped to visual input tokens through the connection module.

Following the configuration of VoCoT, each object is represented by a visually grounded format: “{textual description} [c] {coordinates} [/c] {visual representation}”, e.g., “dog [c] 0.27, 0.08, 0.92, 0.81 [/c] V_{dog} ”. “[c]” and “[/c]” are special tokens denoting the beginning and ending of coordinates. We use bounding boxes $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ as coordinates, x and y are normalized between 0 and 1 w.r.t to the image size. Coordinates are treated as text, undergoing tokenization and embedding.

In addition to text and coordinates, visual tokens

of objects, such as V_{dog} , are supplemented to help the model reference the corresponding visual information in images. The visual tokens of objects are obtained based on the coordinates and image tokens through the RefBind mechanism. Once the end of coordinates token, “[/c]”, is detected in the input or generated, RefBind is activated to obtain the visual object tokens based on the coordinate between “[c]” and “[/c]”. The obtained object tokens are further appended after the “[/c]” token.

RefBind A straightforward method for representing objects is to crop the corresponding regions in the image and encode them with the ViT. However, this method introduces additional computational costs and loses the contextual information of the complete image. For regions with very few pixels, representing objects with sub-images would introduce redundant information. To tackle with above issues, we propose the RefBind mechanism.

RefBind (short for “Reffering Bind”) is conceptually illustrated in Figure 3. Inspired by the RoI-pooling method in Fast-RCNN (Girshick, 2015), given a bounding box and the encoded 2D grid features of the entire image, we can efficiently index the patches in which the target object appears. The features of these patches are flattened into a sequence that represents the object. RefBind relies solely on indexing operations without additional computation. Additionally, the object representation obtained by RefBind inherently preserves contextual information within the whole image.

3.2 Training

The training of VolCano undergoes three stages:

Stage 1: Alignment Pre-training The first stage aims to align visual representations with the LLM

backbone. We utilize image-caption pairs from LLaVA-Pretrain (Liu et al., 2023b). Only the parameters in the connection module are updated.

Stage 2: Multi-modal Interleaved Pre-training

Following the alignment stage, the model is trained to adapt to multi-modal interleaved sequences and visually grounded object representations with Ref-Bind. Three types of sequences are considered: (1) **Image-caption pair** constitutes the simplest form of multi-modal sequences. We utilize ALLaVA-Caption (Chen et al., 2024) which provides detailed descriptions. (2) **Multi-modal document** includes multiple image-text pairs in a sequence. Based on the relevance between images and text, we filter a subset of documents from MMC4 (Zhu et al., 2024). (3) **Grounded image caption** further annotates the coordinates of objects in the caption. We extend the object representations to the visually grounded format consistent with VoCoT. Flickr30K Entities (Plummer et al., 2015) and a subset of GRIT (Peng et al., 2023) are adopted. Both the connection module and the LLM backbone are trained to model the multi-modal sequences.

Stage 3: Instruction Tuning The pre-trained model is further fine-tuned to follow instructions in multi-modal contexts and perform multi-step reasoning with VoCoT. We supplement the constructed VoCoT-Instruct-80K and referring expression data (Kazemzadeh et al., 2014; Chen et al., 2023b) to the existing non-CoT-form visual instruction data (Liu et al., 2023b). We update the LLM backbone and connection module in this stage.

Table 1 summarizes the data mixtures used during the three training stages of VolCano. For more details, please refer to Section 4.1 and Appendix A.3.

4 Experiments

4.1 Experiment Settings

Implementation Details We build VolCano with the pre-trained ViT-L/14 CLIP (Radford et al., 2021) visual encoder and Mistral-7B (Jiang et al., 2023) as the baseline backbone. In addition, we explore the impact of a more powerful LLM backbone in our framework, constructing VolCano_{Q2} based on Qwen2-7B (Yang et al., 2024). Detailed parameter settings are provided in Appendix A.3. To save resources, we merely evaluate VolCano_{Q2} in the main experiments and primarily focus on the Mistral-based VolCano in further analysis.

Stages	Data Type	Source	Size
Stage 1	Image-Caption	LLaVA	558k
	Image-Caption	ALLaVA	695k
Stage 2	Grounded Image-Caption	GRIT	756k
		Flickr30k	148k
	Multimodal Document	MMC4	890k
Stage 3	Visual Instruction	LLaVA	612k
	Referring Expression	Shikra-RD	6k
		RefCOCO	42k×3
		g-RefCOCO	79k
VoCoT	This Work	80k	

Table 1: The training data mixtures used by VolCano.

Evaluation Benchmarks To validate the effectiveness and versatility of the VoCoT framework, we adopt different tasks across various scenarios for assessment: (1) **General VQA** benchmarks, including GQA (Hudson and Manning, 2019), MMBench (Liu et al., 2023c), and SEED (Li et al., 2023a); (2) **Composite tasks** requiring multi-step analysis and composite capabilities, such as visual spatial reasoning in VSR (Liu et al., 2023a) and EmbSpatial (Du et al., 2024), visual search in V-Star (Wu and Xie, 2023), complex reasoning in CLEVR (Johnson et al., 2017) and Winoground (Thrush et al., 2022), and complex referring expression in CLEVR-Ref (Liu et al., 2019); (3) **Hallucination** benchmarks, including POPE (Li et al., 2023d) and AMBER (Wang et al., 2023b), to evaluate whether VoCoT can mitigate hallucinations. AMBER uses CHAIR (Wang et al., 2023b) as the evaluation metric while accuracy is reported for other datasets. Details on the evaluation processes are provided in Appendix A.4.

Baselines We compare VolCano with existing LMMs with ~7B parameters, as listed in Appendix A.5. For strict comparison, we construct a baseline model, VolCano-SE, which is based on the same architecture as VolCano but without VoCoT-Instruct-80K training data, so it can only perform single-step reasoning. We divide models into two groups for comparison: models based on baseline backbones (LLaMA-1,2, Vicuna, Qwen, and Mistral) and models based on advanced backbones (LLaMA-3 and Qwen2). We focus on models with single-image inputs in the main part. Please refer to Appendix B.1 for comparison and discussion involving models that use multiple additional sub-images as inputs for resolution enhancement.

Model		General VQA				Composite Tasks						Hallucination	
Method	Res.	#VP	GQA	MMB ^{Dev}	Seed ^I	VSR	EmbSpa.	CLEVR	V-Star	Wino ^{txt}	C-Ref	POPE ^A	AMB [↓]
Models based on baseline LLM backbones													
InstructBLIP-7B	224 ²	1.3B	49.20	36.00	-	52.10	33.41	-	34.02	-	-	72.10	8.80
Shikra-7B	224 ²	0.3B	-	58.80	-	-	34.75	-	-	-	-	83.10	-
mPLUG-Owl2-7B	448 ²	0.3B	56.10	<u>64.50</u>	59.99	-	36.72	43.22	36.12	63.38	-	-	10.60
MiniGPT-v2-7B	448 ²	1.3B	60.10	55.14	51.50	62.90	<u>43.85</u>	46.23	33.19	62.00	24.90	80.50	-
Qwen-VL-Chat	448 ²	1.9B	57.50	60.60	64.70	-	38.68	<u>53.20</u>	45.80	-	22.35	84.70	<u>5.50</u>
LLaVA1.5-7B	336 ²	0.3B	62.00	64.30	53.80	64.24	42.43	43.73	48.68	55.31	6.70	84.50	7.80
VILA-7B	336 ²	0.3B	62.30	61.50	60.40	<u>66.02</u>	38.05	47.60	46.22	<u>66.37</u>	-	84.50	10.50
VisCoT-7B	336 ²	0.3B	<u>63.00</u>	63.82	63.23	-	37.01	53.15	61.76	56.40	<u>32.05</u>	<u>86.10</u>	7.20
VolCano-SE	336 ²	0.3B	59.91	61.15	54.15	63.42	36.14	51.70	44.96	64.00	21.70	84.50	6.70
VolCano	336 ²	0.3B	64.40	68.10	<u>64.50</u>	67.18	58.29	56.17	<u>58.40</u>	68.37	33.95	86.50	4.60
Models based on advanced LLM backbones													
VILA1.5-8B	384 ²	0.4B	63.50	64.38	64.41	53.76	54.95	<u>55.22</u>	<u>58.74</u>	66.00	-	84.90	8.50
Bunny-8B V1.0	384 ²	0.4B	<u>64.00</u>	<u>70.86</u>	67.59	65.71	53.54	54.47	58.32	<u>68.50</u>	-	<u>86.40</u>	<u>7.40</u>
VolCano-SE _{Q2}	336 ²	0.3B	62.23	66.87	64.51	<u>69.37</u>	<u>55.19</u>	51.58	56.30	66.63	<u>23.90</u>	85.20	8.00
VolCano _{Q2}	336 ²	0.3B	64.60	71.61	<u>66.95</u>	74.22	59.86	56.78	62.81	68.78	34.00	86.60	4.40
GPT-4V	2048 ²	-	-	75.80	71.60	68.24	36.07	51.90	55.00	83.75	-	82.00	4.60

Table 2: **Comparison on 11 benchmarks.** Res. and #VP respectively denote the input image resolution and the number of parameters in visual encoder. MMB^{Dev}, Seed^I, EmbSpa., Wino^{txt}, C-Ref, POPE^A and AMB represent MMbench-DEV, SEED-Image, Embspatial, the reformulated Winoground, CLEVR-Ref, POPE-adversarial, and AMBER, respectively. ↓ indicates the lower metric is preferred. For each dataset, the best result in each group is highlighted in **bold** while the runner-up is underlined. Except GQA, all results are evaluated in a zero-shot manner.

Method	Obj-Format	Seed ^I	EmbSpa.	CLEVR	V-Star	VSR	Wino ^{txt}	POPE ^A	AMB ^{cover}	AMB ^{chair} ↓
Zero-Shot CoT	< T >	56.79	52.47	51.70	45.32	57.20	65.00	67.50	52.20	6.70
Text CoT	< T >	63.36	59.20	49.60	47.90	68.49	65.75	84.63	49.30	5.50
Coor. CoT	< T, C >	<u>64.32</u>	<u>58.59</u>	<u>54.42</u>	<u>53.78</u>	66.86	<u>65.87</u>	85.47	47.80	4.30
Sub-Img CoT	< T, C, S >	61.29	54.10	51.85	63.45	66.23	58.87	<u>85.77</u>	47.80	<u>4.60</u>
VoCoT	< T, C, R >	64.50	58.29	56.17	<u>58.40</u>	<u>67.18</u>	68.37	86.50	<u>51.00</u>	<u>4.60</u>

Table 3: **Comparison between different CoT formats.** The “Obj-Format” column indicates the representation format of objects. T, C, S, and R are short for texts, coordinates, sub-images, and RefBind representations.

4.2 Main Results

Table 2 presents a thorough evaluation of existing LMMs. Several insights can be gleaned: (1) By comparing VolCano and VolCano-SE, it demonstrates that VoCoT effectively mitigates hallucinations and brings consistent improvement across all benchmarks. Section 4.3 delves deeper into how VoCoT contributes to reliable and visually grounded reasoning. (2) Across different datasets, VolCano and VolCano_{Q2} achieve the best or second-best results within their respective group, where the advantages are more pronounced in composite tasks. On benchmarks like CLEVR and EmbSpatial, VolCano with a limited scale even outperforms powerful GPT-4V. (3) Furthermore, we compare two multi-modal CoT methods: VisCoT and VoCoT. VisCoT (Shao et al., 2024) designs a simple two-step reasoning process: first searching for a single relevant region and then answering based on the detected region. Experimental results imply that VisCoT merely performs better on V-Star because

the questions in V-Star perfectly align with the two-step search logic of VisCoT. However, VisCoT falls short in other complex scenarios that involve interaction between multiple objects, indicating that VoCoT is a more generalizable format of multi-modal CoT. Overall, the experimental results validate the effectiveness of VoCoT-based multi-step reasoning in various scenarios. In addition, we find that VoCoT could seamlessly generalize to other scenarios including scene-text-centric tasks, please refer to the results and analysis in Appendix B.2.

4.3 Comparing CoTs in Different Formats

We validate the effectiveness of the VoCoT format by comparing it with different CoT formats: (1) Zero-Shot CoT directly prompts VolCano-SE to think step-by-step without training; (2) Text CoT represents objects with only text descriptions; (3) Coor. CoT augments Text CoT with coordinates; and (4) Sub-Img CoT encodes sub-images as representations of objects rather than using RefBind.

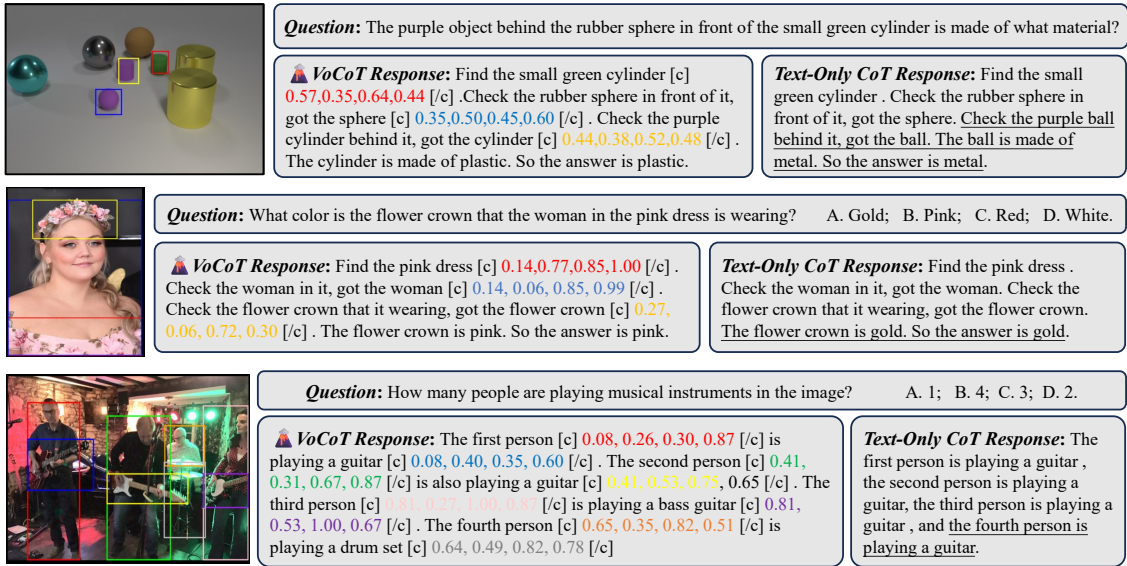


Figure 4: Qualitative analysis to compare VoCoT and text-only CoT. Hallucinations are underlined.

Stage 2	Stage 3			Seed ^l	EmbSpa.	CLEVR	V-Star	VSR	Wino ^{txt}	POPE ^A	AMB ^{cover}	AMB ^{chair} ↓
	Type 1	Type 2	Type 3									
✓	✓			63.63	47.78	<u>56.07</u>	57.14	<u>68.90</u>	66.00	84.80	33.70	1.80
✓		✓	✓	65.45	<u>58.21</u>	54.93	61.34	66.85	65.00	80.16	<u>48.90</u>	5.00
✓	✓	✓		64.20	57.24	54.17	53.36	68.98	<u>66.87</u>	86.50	<u>48.90</u>	<u>4.40</u>
✓	✓	✓	✓	<u>64.50</u>	58.29	56.17	<u>58.40</u>	67.18	68.37	86.50	51.00	4.60
	✓	✓	✓	61.62	57.22	49.18	53.36	67.13	63.62	<u>85.60</u>	46.90	5.50

Table 4: Ablation of VoCoT-formatted data on stage 2 and stage 3 training process.

Table 3 lists the results. Firstly, the zero-shot multi-step reasoning capability of VolCano-SE is limited. It is likely to exhibit hallucinations, highlighting the necessity to construct visual CoT tuning data. Secondly, CoT expressed only in texts is also affected by hallucinations, handling spatial reasoning well where each type of object appears only once, but failing to manage more complex scenarios. Thirdly, introducing coordinates grounds the thoughts to visual signals, mitigating hallucination and improving performance across various tasks. Furthermore, representations obtained by RefBind effectively help the model to utilize visual signals of objects. In contrast, the performance of Sub-Img CoT is overall inferior to that of Coord. CoT, which supports our claim in Section 3.1: simply encoding each object as a sub-image may introduce redundant information and degrade the performance.

Besides quantitative results, we present cases in Figure 4. We observe that text-only CoT is limited in terms of: (i) It may fail to accurately find/locate target object (Case 1). (ii) It is unable to leverage object-level visual information for inferring object attributes, as VoCoT does through RefBind (Case

3). (iii) It cannot resolve ambiguity between multiple objects (Case 4). (iv) Lack of interpretability. In general, it is crucial to ground the reasoning process to the visual information and VoCoT is the most suitable format.

4.4 Ablation on the Constructed Dataset

In Table 4, we explore the role of three types of data in VoCoT-Instruct-80K. The results implies that: (1) Type 1, the GQA-based data, is precise but limited in terms of diversity. Models trained solely on Type 1 data produces the fewest hallucinations but struggle to handle diverse questions. (2) Type 2 and 3 data effectively help the model generalize across various instructions. Nevertheless, totally removing Type 1 data will increase the risk of hallucinations. (3) Introducing multi-modal interleaved data in Stage 2 leads to a significant improvement. In summary, interleaved pre-training data and three types of VoCoT data should be jointly utilized.

4.5 Further Analysis

VoCoT enhances performance in complex questions Figure 5 compares the performance of

Analyzer	VolCano _V			VolCano				VolCano _{Q2}	GPT-4V
Judger	VolCano _V	Vicuna-1.5	Mistral	VolCano	Vicuna-1.5	Mistral	GPT-4		
Accuracy (%)	63.5	64.5	67.2	67.2	65.1	67.8	73.8	74.2	68.2

Table 5: Performance with different analyzers and judges on the VSR benchmark.

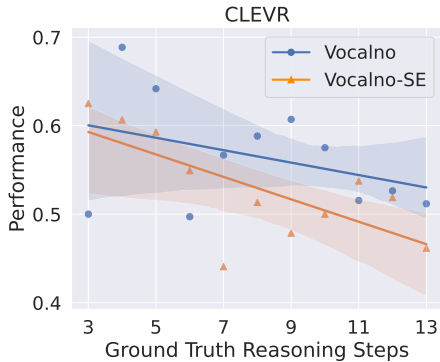


Figure 5: Relationship between performance and the number of reasoning steps required by the questions.

VolCano-SE and VolCano on questions of varying difficulty in CLEVR. The fitted curves and confidence intervals imply that as the number of required reasoning steps increases, the advantage of multi-step reasoning becomes more pronounced.

Disentangling Multi-Modal Reasoning Our preliminary study finds that sometimes VolCano generates reasonable reasoning paths but fail to infer the correct answer. Therefore, we split the reasoning process into two sub-processes: analysis and judgment, where the former constructs reasoning paths and the latter provides conclusions. We conduct experiments to combine different analyzers and judges on the VSR benchmark, where each object category corresponds to a single object in the image, allowing us to use text-only LLMs to judge based on the paths analyzed by VolCano. In Table 5, VolCano_V represents the Vicuna-based VolCano to explore the impact of LLM backbones. The results indicate that the judge plays a important role. The path analyzed by VolCano help GPT-4 make better decisions (73.8%) than GPT-4V (68.2%). However, the overall capability of VolCano is upper-bounded by the judgement ability of its LLM backbone. Comparison between VolCano_V, VolCano, and VolCano_{Q2} further reveals the potential of applying VoCoT on stronger LLM backbones.

Case Study Examples in Figure 6 show that VolCano provides a visually grounded description with no hallucinations in AMBER. In CLEVR, VolCano infers effective reasoning paths towards the answer. See Appendix B.8 for more cases in other datasets.

5 Related Works

5.1 Large Multi-Modal Models

Architecture of LMMs A vast amount of research has emerged, focusing on adapting LLMs to handle multi-modal tasks. Initially, researchers treat LLMs as intelligent agents capable of using various tools. They train or prompt LLMs to invoke fundamental vision models, enabling them to complete multi-modal tasks such as captioning and VQA (Wu et al., 2023; Yang et al., 2023c). Recent methods directly align the visual and textual representations in a unified backbone. To achieve this, various connection modules are designed, including MLP (Liu et al., 2024b, 2023b), Q-Former (Li et al., 2023b; Dai et al., 2023), and cross-attention layers (Alayrac et al., 2022; Wang et al., 2023c).

Visual Instruction Data Construction To enable LMMs to follow instructions in multi-modal contexts, a line of research has focused on converting existing academic datasets into instruction data (Dai et al., 2023; Li et al., 2023c). In addition, researchers also use powerful tools like GPT-4V to assist in bootstrapping and generating high-quality data (Zhu et al., 2023; Liu et al., 2024b). Further efforts are devoted towards improving both the quality and scale of the generated data (Chen et al., 2024; Zhao et al., 2023; Wang et al., 2023a).

Visual Grounding in LMMs Beyond text descriptions, some researchers further empower LMMs to understand and extract fine-grained visual information through visually grounded representations like coordinates (Peng et al., 2023; Chen et al., 2023b), masks (Yuan et al., 2024) and visual prompts (Cai et al., 2024). However, most of these LMMs still rely on single-step and text-only reasoning, failing to effectively leverage fine-grained information during the reasoning process.

5.2 Multi-Step Reasoning

CoT in LLMs Chain of Thoughts (CoT) is a series of prompting techniques designed to facilitate LLMs in addressing complex problems by guiding them through intermediate steps. CoT is first proposed through in-context learning (Wei et al., 2022), followed by the introduction of zero-shot CoT (Kojima et al., 2022), Auto-CoT (Zhang et al.,

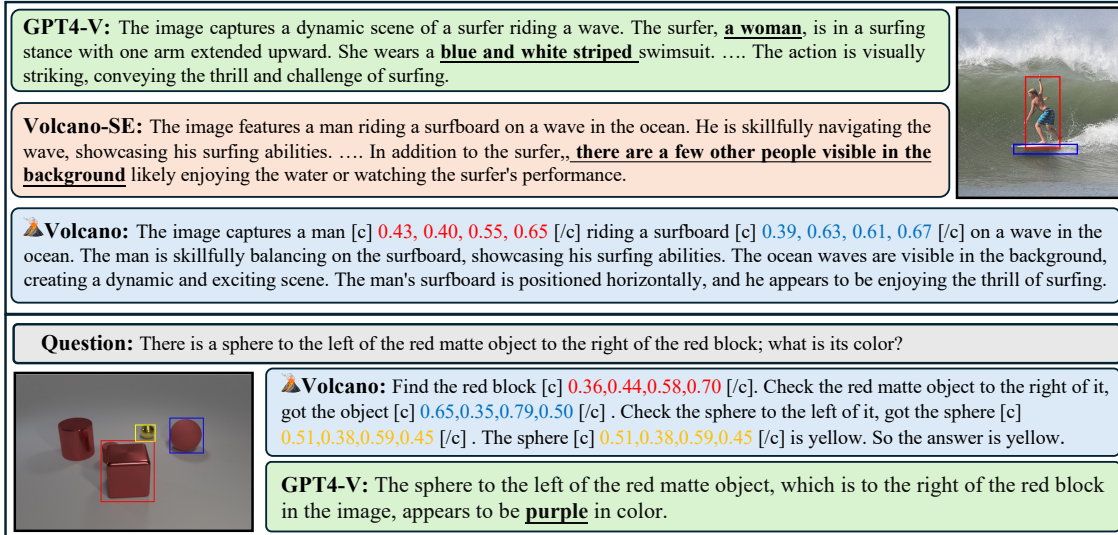


Figure 6: Qualitative analysis with cases from AMBER and CLEVR. **Hallucinations** are highlighted.

2022) and self-consistency (Wang et al., 2023d). Subsequently, CoT are extended to more complex formats (Yao et al., 2024; Besta et al., 2024).

Visually Enhanced Reasoning To address complex multi-modal problems, Shikra (Chen et al., 2023b) initially explores the potential of applying CoT to specific tasks with LMMs. SoM (Yang et al., 2023a) and Scaffolding (Lei et al., 2024) respectively incorporate segmentation maps and dot grids in images to assist LMMs in reasoning, but such information can only be utilized by proprietary models like GPT-4V. The most related work is VisCoT (Shao et al., 2024), which designs a two-step CoT: first searching for a relevant region and then answering based on the additional region information. This method is effective but cannot model complex multi-step reasoning. Overall, CoT has not been comprehensively explored in LMMs and there lack appropriate reasoning formats that could be generalized to various scenarios and tasks.

6 Conclusion

In this paper, we introduce VoCoT, a visually-grounded and object-centric chain of thoughts format to assist LMMs in multi-step reasoning. We also curate a VoCoT-formatted dataset from existing resources to train LMMs to learn reasoning with VoCoT. Building on this, we develop VolCano, a model capable of multi-step reasoning using the VoCoT format. Comprehensive experimental results demonstrate the effectiveness of our approach.

Limitations

Our work, as an early exploration of CoT techniques in large multi-modal models, is limited in the following aspects. (1) Currently, VoCoT is designed for single-image context and not applicable to multi-image inputs like videos and image sequences. Additional special tokens or marks can be introduced to extend VoCoT to a two-step grounding for multiple images. Each object is first grounded to a specific image and then localized to a region within that image. We will explore such mechanisms in our future work. (2) The construction of VoCoT-formatted dataset is limited by the cost of calling proprietary models and can not effectively scale up. In future work, we will explore methods to reduce the cost of data construction, including using smaller or open-source models, collecting and converting more finely annotated data (such as DocVQA) in a manner similar to GQA, and simulating and generating data based on specific needs, similar to CLEVR. (3) The presented VolCano model is currently limited with respect to 7B-sized models due to the lack of computational resources. As implied by the experimental results in Section 4.5, we hope to demonstrate the potential of applying VoCoT to larger and stronger backbones as explored in textual CoT techniques.

Ethical Statement

The presented VoCoT-Instruct-80K dataset is sourced from open-source datasets including GQA (Hudson and Manning, 2019), LLaVA-Instruct (Liu et al., 2024b), and LVIS (Gupta et al., 2019). We carefully follow the license to use these

datasets and ensure that they are applicable for research purposes. The original datasets have been widely adopted by relevant researchers and ensure no risk of privacy leakage or harmful information. Furthermore, during the data collection and construction, we perform balanced sampling based on the distribution of object categories to alleviate distribution bias. As mentioned in Appendix A.1, we also conduct human-in-the-loop quality control to ensure the final dataset has correct information without ethical issues. Please refer to Appendix B.9 for the detailed discussion. Currently, the presented models and dataset focus on English, we hope to expand to other languages in the future. Our work and artifacts are designed with the principle of universality and fairness, without any preference for specific demographic groups.

Acknowledgment

The work is supported by National Key R&D Program of China (Grant Nos. 2023YFF1204800) and National Natural Science Foundation of China (Grant Nos. 62176058). The project’s computational resources are supported by CFFF platform of Fudan University.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, pages 8948–8957.

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. *NIPS*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12914–12923.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

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

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv:2306.15195*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

R Girshick. 2015. Fast r-cnn. *arXiv preprint arXiv:1504.08083*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *Preprint*, arXiv:2402.11530.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. 2023. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.
- 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.
- Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024. *Scaffolding coordinates to promote vision-language coordination in large multi-modal models*. *Preprint*, arXiv:2402.12058.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*.
- Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. *M³it: A large-scale dataset towards multi-modal multilingual instruction tuning*. *Preprint*, arXiv:2306.04387.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*.
- Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, et al. 2023e. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023f. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv:2311.06607*.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll ar, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *TACL*, 11:635–651.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv:2403.05525*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large

- language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209.
- OpenAI. 2023a. [Chatgpt \(august 3 version\)](#).
- OpenAI. 2023b. [Gpt-4 technical report](#). *arXiv:2303.08774*.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shao-han Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- 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. In *ICML*, pages 8748–8763. PMLR.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.
- 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 *CVPR*, 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:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.
- Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv:1601.07140*.
- Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. [To see is to believe: Prompting gpt-4v for better visual instruction tuning](#). *Preprint*, arXiv:2311.07574.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv:2311.07397*.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023c. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023d. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- 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.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. [Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v](#). *Preprint*, arXiv:2310.11441.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of llms: Preliminary explorations

with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023c. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. [mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration](#). *Preprint*, arXiv:2311.04257.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

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

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. [Automatic chain of thought prompting in large language models](#). *Preprint*, arXiv:2210.03493.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NeurIPS*, 36.

A Supplementary Details

A.1 Data Construction Details

Construction Methods As for the Type 1 data construction, we utilize a rule-based conversion method, the mapping rules used are listed in Table 16. In terms of Type 2 and Type 3 data, the prompts for GPT-4V are respectively shown in Table 17 and 18. We use in-context learning to let GPT-4V generate thought with multi-step reasoning path in VoCoT format.

Quality Control For Type 1 data generated by rule based mapping, the quality is controlled by the original source, namely GQA. We do not perform additional quality control.

For Type 2 and Type 3 data generated by GPT-4V, we first perform balanced sampling for images to achieve a balanced distribution of objects included. Secondly, we manually sample and check 200 data samples and find that initially constructed data suffers issues like uneven question types and incorrect object information. The first issue can be addressed by including well-designed in-context samples. The second issue is mainly caused by the potential incomplete and incorrect labels in LVIS. We find if GPT-4V think the information we provide is not enough to generate correct reasoning path, the response will contains error messages. All error message have patterns in common which may contains the following phrases: “From the object information provided”, “provided object information”, “From the bounding boxes provided”, and so on. We remove the samples containing these patterns, leaving 8k samples after the filtering. Ultimately, by manually checking the constructed dataset, we do not observe any bias issues and achieve a 98% pass rate on the presented dataset.

A.2 Training Data Details

We present our data mixture details in Table 8. In Stage 1, we use LLaVA-pretrain (Liu et al., 2023b) dataset for projector alignment which contains 558k image caption pairs. In Stage 2, to better adapt to multi-modal interleaved sequences and visually grounded object representations with Refbind, we mix three types of data: (1) multi-modal documents, (2) grounded image captions, and (3) high-quality image captions. Multimodal documents data is sample from MMC4 (Zhu et al., 2024) by choosing which average similarity score between image and sentence before each image

is larger than 0.3. Each multimodal document have multiple image, we remove samples with more than 6 images. Grounded image captions is from GRIT (Peng et al., 2023) and Flickr30K Entities (Plummer et al., 2015). For GRIT, we filter samples with clip score is larger than 0.35. We also use high quality image caption from ALLaVA (Chen et al., 2024) which is generated by GPT-4V and provide details description. In Stage 3, we remove samples from LLaVA-Instruct that meet two criteria: sourced from RefCOCO, and sourced from VG where the object sub-image size is less 50, the reason is that we find extremely small regions in VG are probably with low quality.

A.3 Model & Training Details

The hyper-parameters in each stage are in Table 7. The learning rate setup mainly follows that of LLaVA (Liu et al., 2024b). In stage 1, a large learning rate is used to update the connection module, aiming to quickly align the cross-modal representations. In the latter stages, a small learning rare is adopted to carefully fine-tune the backbone. Following Kosmos2 (Peng et al., 2023), we introduce a special token “<grounding>” at the beginning of the sequence to control whether to require VolCano to produce visually grounded description.

Also notice that VolCano is able to perform both single-step and multi-step reasoning. Following (Kojima et al., 2022), we introduce a prompt “Answer the question and include the reasoning proess. Locate key objects and provide bounding boxes in your thoughts.” as a trigger to tell the model whether to use VoCoT or not during the generation process.

A.4 Evaluation Details

In this section, we introduce the details in the evaluation procedure.

A.4.1 Benchmark Details

General VQA Benchmarks In GQA, we utilize the “testdev_balanced” split following (Liu et al., 2024b). As for MMBench, we adopt the “DEV” split for the evaluation efficiency. In terms of SEED, we only consider the subset that the visual inputs are images. For GQA, we append a prompt “Please answer in a word or short phrase” to require models produce concise outputs.

Spatial Reasoning Benchmarks For VSR, we utilize the unseen test split for zero-shot evaluation.

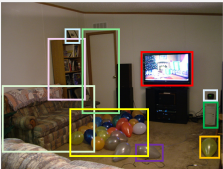
<p>Type 1 (GQA Source) Rule Input [Object Info]: Shelf: [0.224, 0.219, 0.386, 0.592], Door: [0.394, 0.176, 0.524, 0.645] [Question]: What is common to the shelf and the door? [SemanticStr]: select: shelf -> select: door -> common: [0, 1] [FullAnswer]: The material, both the shelf and the door are wooden. [Answer]: material.</p> <p>Type 1 (GQA Source) Response [Thought]: Find the shelf [0.224, 0.219, 0.386, 0.592]. Find the door [0.394, 0.176, 0.524, 0.645] . The question ask the common attribute of the two objects. The material, both shelf [0.224, 0.219, 0.386, 0.592] and door [0.394, 0.176, 0.524, 0.645] are wooden. So answer is material.</p>	
<p>Type 2 (VQA-Based Source) GPT-4V Input [Object Info]: TV: [0.78, 0.84, 0.97, 0.98], Tie: [0.72, 0.12, 0.90, 0.60], Toothbrush: [0.15, 0.37, 0.56, 0.49], Person: [0.0, 0.13, 0.81, 1.0] [Question]: What might be the purpose behind the woman’s action? [Answer]: The purpose behind her actions may be for entertainment.</p> <p>Type 2 (VQA-Based Source) Response [Thought]: The woman [0.0, 0.13, 0.81, 1.0] is engaging in an unusual activity by using a toothbrush [0.15, 0.37, 0.56, 0.49] that is significantly larger than a standard one and wearing a tie [0.38, 0.6, 0.59, 0.99] that also appears to be oversized. . . . It’s plausible that her actions are meant to entertain or educate audiences.</p>	
<p>Type 3 (Image-Only Source) GPT-4V Input [Object Info]: Ferris wheel: [0.09, 0.40, 0.14, 0.54] , Building: [0.40, 0.36, 0.58, 0.53], Statue: [0.72, 0.12, 0.90, 0.60], Dock: [0.00, 0.56, 0.33, 0.67], Stair: [0.01, 0.69, 1.00, 1.00],</p> <p>Type 3 (Image-Only Source) Response [Question]: Where is this location? [Referring of Thought]: First, The statue [0.72, 0.12, 0.90, 0.60] is Merlion Statue, a mythical creature with the head of a lion and the body of a fish, which is a mascot of Singapore. Secondly, the building [0.40, 0.36, 0.58, 0.53] . . . Based on the distinctive architecture of the landmarks, this image is taken in Singapore. [Answer]: Singapore.</p>	

Table 6: Examples to illustrate the construction of VoCoT-formatted data from three data sources. Type 1 data are obtained by rules, while Type 2 and Type 3 data are obtained by leveraging GPT-4V.

For each sample, a description is provided and the model is required to distinguish if the claims is supported by the image. We use the prompt “Is there a event {description} in the image?” for this dataset. With respect to EmbSpatial, we use the test split for assessment.

Hallucination Benchmarks For POPE, we consider the adversarial subset since it is the most challenging split. In AMBER, we leverage the generative task which asks the model to describe the image. All prompts are adopted from the original datasets with a yes-or-no instruction for POPE.

Benchmarks for Composite Tasks For CLEVR, we utilize the val split. Because the original CLEVR validation set is too large, we categorize the data the into six types based on the question type: count, yes/no, shape, material, size, and color. We sample 1k questions from each category as test samples and construct a multiple-choice candidate set based on the feasible answers in the dataset. We will also open-source this subset. For Winoground,

we utilize the test set and consider it as a caption selection multiple-choice question, the prompt is designed as “Please describe the image.”. As for V-Star, we directly use the V-Star benchmark. Regarding CLEVR-Ref, which is a referring expression task with relatively complex queries, we use the provided set for evaluation. We design a prompt as “Can you locate {phrase} in the image?” where “{phrase}” is the target query.

Please see Table 10 and Table 9 for the splits and scales of benchmarks used in this paper.

A.4.2 Evaluation Methods

All evaluation benchmarks adopted in this paper can be divided into three categories based on the task formulation: multiple-choice questions, open-ended generation, referring expression.

Multiple-Choice Question For multiple-choice questions, we utilize the likelihood-based evaluation method, which is also known as the perplexity-based method. These methods are widely adopted in evaluating LMMs (Li et al., 2023a,b; Dai et al.,

Configuration	Alignment	Multi-modal Interleaved	Instruction Tuning
Visual Encoder	OpenAI-CLIP ViT-L/14	OpenAI-CLIP ViT-L/14	OpenAI-CLIP ViT-L/14
Backbone Init	Mistral-Chat-v0.2-7B	Stage1	Stage2
Optimizer	AdamW	AdamW	AdamW
Optimizer Hyperparameters	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e^{-6}$	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e^{-6}$	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e^{-6}$
Global batch size	256	128	128
Peak learning rate of LLM	1e-3	1e-5	1e-5
Learning rate schedule	Cosine	Cosine	Cosine
Training Epochs	1	1	1
Warm-up ratio	0.03	0	0
Weight decay	0.0	0.0	0.0
Gradient clipping	1.0	1.0	1.0
Input image resolution	336 * 336	336 * 336	336 * 336
Input sequence to LLM	2048	3072	3072
Numerical precision	bfloat16	bfloat16	bfloat16
GPU Usage	8 NVIDIA A100	8 NVIDIA A100	8 NVIDIA A100
Training Time	12h	48h	30h

Table 7: The detailed training hyper-parameters of VolCano. Except for the backbones for initialization, VolCano_{Q2} follows the same hyper-parameters.

Stages	Data Type	Source	Size
Stage 1	Image-Caption	LLaVA-Pretrain (Liu et al., 2024b, 2023b)	558k
Stage 2	Image-Caption	ALLaVA-Caption (Chen et al., 2024)	695k
	Grounded Image Caption	GRIT (Peng et al., 2023)	756k
		Flickr30k-Entities (Plummer et al., 2015)	148k
	Multimodal Document	MMC4 (Zhu et al., 2024)	890k
Stage 3	Visual Instruction	LLaVA (Liu et al., 2023b)	612k
	Referring Expression	Shikra-RD (Chen et al., 2023b)	6k
		RefCOCO (Kazemzadeh et al., 2014)	42k
		RefCOCO+ (Kazemzadeh et al., 2014)	42k
		RefCOCOG (Kazemzadeh et al., 2014)	42k
		g-RefCOCO (He et al., 2023)	79k
VoCoT	This Work	80k	

Table 8: The data mixture used in the three training stages of VolCano.

2023; Li et al., 2023e). The key idea is to select the option with the highest generated likelihood, please refer to these papers for the detail. If VoCoT is utilized, the likelihood is computed based on the question, image, and the generated reasoning path.

Open-Ended Generation For GQA, we use the evaluation script provided by LLaVA (Liu et al., 2023b) for a fair comparison. As for VSR and POPE, we require the model to answer in yes and no, enabling us to evaluate the correctness with exact match. With respect to AMBER, we use the official evaluation method to assess the hallucinations in the generated descriptions.

Referring Expression We first extract the predicted boxes from the outputs based on rules, then calculate the IoU between the ground truth box and the predicted box. If the IoU is larger than 0.5,

it is considered as a correct prediction following (Kazemzadeh et al., 2014).

Further Analysis Setup In the reasoning capability assessment part in Section 4.5, to leverage LLM as the judge model. We utilize a prompt “There is a image, {reasoning path}, please determine whether {description}, please answer yes or no.”, where the reasoning path are generated by the analyzer (with the coordinates and visual information removed), descriptions are the target description in VSR. If the model chooses not to predict, we consider the prediction as “no”.

A.5 Introduction to Baseline Models

We compare VolCano to several existing SOTA open-source LMMs, including BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023),

Category	General VQA			Spatial Reasoning		Hallucination	
Dataset	GQA	MMBench	SEED	VSR	EmbSpat.	POPE	AMBER
Split	testdev_balanced	DEV	Image	test unseen	test	adversarial	generative
Size	12578	4329	14233	1222	3625	3000	1004

Table 9: Information of Evaluation Benchmarks.

Category	Composite Tasks			Referring Expression
Dataset	V-Star	Wino	CLEVR	CLEVR-Ref
Split	-	test	val	-
Size	238	800	6000	2000

Table 10: Supplementation of Table 9.

Shikra (Chen et al., 2023b), mPLUG-Owl2 (Ye et al., 2023), MiniGPT-v2 (Chen et al., 2023a), Qwen-VL-Chat (Bai et al., 2023), VILA (Lin et al., 2024), LLaVA-1.5 (Liu et al., 2023b), and the most related VisCOT (Shao et al., 2024). These models are based on baseline LLM backbones released in 2023, including LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023), Mistral (Jiang et al., 2023), and Qwen (Bai et al., 2023). For models based on recently proposed advanced backbones like LLaMA-3 (Dubey et al., 2024) and Qwen2 (Yang et al., 2024), we compare VolCano_{Q2} with Bunny (He et al., 2024) and VILA-1.5 (Lin et al., 2024). The models listed before take a single image as input, which is consistent with VolCano and VolCano_{Q2} for a fair comparison. Additionally, we include another series of SOTA LMMs that enhance input resolution by splitting a single image into multiple sub-images: LLaVA-1.6 (Liu et al., 2024a), Deepseek-VL (Lu et al., 2024), and Monkey (Li et al., 2023f). As for GPT-4V, examples in Figure 1 and all results are obtained by calling openai API using the “GPT-4V” model between 2024/05/18 to 2024/05/30.

We only consider zero-shot performance in Table 2, except for GQA. If a model has been trained on a specific evaluation benchmark, we do not report the corresponding evaluation results. For example, in Table 2, the result of VisCOT on VSR is omitted because it uses the corresponding training data. For certain models, including Bunny and VILA, due to the lack of clear evaluation details, we re-evaluate their performance in the same setting to make a fair comparison with VolCano.

B Supplementary Results and Discussion

B.1 Comparison between VolCano and High-Resolution models

In Table 2 we compare models with single-image and relatively low-resolution inputs. Comparing VolCano with LMMs that enhance input resolution by introducing multiple-image inputs, we observe that these methods primarily improve the performance in general VQA and V-Star, as V-Star provides high-resolution input images (Wu and Xie, 2023). However, in tasks that require complex reasoning, the improvement brought by higher resolutions becomes less significant. VolCano either exceeds or perform comparably with these models in such tasks, indicating the superiority of introducing multi-step reasoning over enriching the input information in these scenarios.

Notice that the RefBind mechanism introduced in Section 3 can be directly extended to fit multiple split sub-images by mapping the predicted coordinates to patches from different sub-images. We leave exploring combining these two vertical research directions—enhancing input resolution and introducing multi-step reasoning—as future work.

B.2 Performance in Additional Benchmarks

The benchmarks presented in the main text primarily focus on object-centric scenarios, which align with the design of VoCoT. Recently, a line of research develops LMMs to handle scene-text-oriented scenarios including document understanding and chart information extraction. In this section, we explore whether VoCoT can adapt to other scenarios by conducting experiments on additional benchmarks: TextVQA (Singh et al., 2019), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), and DocVQA (Mathew et al., 2021) for scene-text-oriented tasks; MMMU (Yue et al., 2023) and MathVista (Lu et al., 2023) for reasoning based on knowledge and scene texts; MMVet (Yu et al., 2023) and MMT (Ying et al., 2024) for instruction-following; COCO caption (Veit et al., 2016) and NoCaps (Agrawal et al., 2019). For im-

Model		General VQA			Spatial Reasoning		Composite Tasks			Hallucination	
Method	Res.	#VP	GQA	MMB ^{Dev}	Seed ^I	VSR	EmbSpa.	CLEVR	V-Star	Wino ^{txt}	POPE ^A AMB [↓]
Models with single-image inputs											
VolCano-SE	336 ²	0.3B	59.91	61.15	54.15	63.42	36.14	51.70	44.96	64.00	84.50 6.70
VolCano	336 ²	0.3B	64.40	68.10	64.50	67.18	58.29	56.17	58.40	68.37	86.50 4.60
Models with multiple-image inputs											
LLaVA1.6-7B	672 ²	0.3B	64.20	68.40	66.15	66.86	56.82	50.35	58.80	64.88	86.90 -
LLaVA1.6-7B _m	672 ²	0.3B	64.80	69.00	67.72	63.77	56.55	51.85	60.08	65.75	86.70 -
Deepseek-VL-7B	1024 ²	0.4B	-	71.32	70.40	67.51	41.77	48.77	62.18	64.88	85.77 -
Monkey	896 ²	1.9B	60.70	61.95	67.58	62.93	32.91	46.33	67.23	68.63	82.57 -
GPT-4V	2048 ²	-	-	75.80	71.60	68.24	26.65	51.90	55.00	83.75	82.00 4.60

Table 11: **Comparison between VolCano and resolution-enhanced models.** The notations follows Table 2. LLaVA1.6-7B_m represents the Mistral-based LLaVA1.6-7B.

Model	Scene-Text-Oriented Tasks				Know. Reasoning		Ins. Following		Captioning	
	TextVQA ^N	AI2D	ChartQA	DocVQA	MMMU	MathVista	MMVet	MMT	COCO	NoCaps
LLaVA1.5-7B	46.1	43.0	14.9	2.9	28.4	26.7	30.5	45.0	94.5	95.6
VolCano-SE	45.1	45.3	19.0	4.6	29.3	27.8	32.1	44.1	81.0	90.6
VolCano	48.9	45.6	19.5	10.6	33.3	27.9	32.9	45.4	100.4	103.5

Table 12: Performance on additional benchmarks. TextVQA^N represents the TextVQA benchmark without providing reference OCR like in LLaVA-1.5 (Liu et al., 2023b). “Know.” and “Ins.” are respectively short for knowledge and instruction. The best performance for each dataset are **bolded**.

age captioning, we report the CIDEr (Vedantam et al., 2015) metric following (Li et al., 2023b) while accuracy is provided for other tasks. For efficiency, we only consider LLaVA-1.5, VolCano-SE, and VolCano for a fair and straightforward comparison to validate the effect of VoCoT.

Scene-Text-Oriented Benchmarks As shown in Table 12, we can find: (1) Quantitatively, VoCoT can generalize to text and chart-centric scenarios and improve the performance, even though VoCoT-Instruct-80K does not include similar data. (2) Qualitatively, we present several samples in Figure 7. VoCoT can treat text blocks and chart areas as “objects”. By locating and analyzing the corresponding regions, VoCoT easily generalizes to such tasks. (3) The cross-domain improvements are very exciting, and the potential of VoCoT in such tasks can be further unleashed by enhancing input resolution and constructing text-centric VoCoT data (with fine-grained annotations in TextVQA, AI2D...), which we leave as our future work.

Knowledge Reasoning and Instruction Following According to results in Table 12, we observe that: (1) Although MMMU and MathVista focus on reasoning involving commonsense, knowledge, and mathematical deduction, the results indicate that the generalization ability of VoCoT can still improve the performance. As shown in Figure 7, the framework of VoCoT to locate and analyze

local objects and regions can generalize to such tasks. (2) Improvements brought by VoCoT in the absence of similar training data demonstrate its potential in such scenario, inspiring us to incorporate both visually grounded information and conceptual knowledge in VoCoT framework in our future work. (3) In general instruction-following datasets, MMT and MMVet, it is shown that VoCoT does not hurt the instruction-following ability and brings improvement, which is consistent with the findings in SEED and MMBench mentioned in our main paper. See Appendix B.3 for further exploration.

Image Captioning Results in Table 12 imply that VoCoT helps VolCano to perceive accurate information and produce visually grounded descriptions, improving the quality of generated captions.

Generally, the results validates that **VoCoT can generalize across various tasks and demonstrates potential for further enhancement.**

B.3 Do VoCoT Damage the Original Capabilities of LMMs?

Another concern about whether the proposed VoCoT framework brings negative effects to the original capabilities of a LMM. Noticing that the trained VolCano can perform both single-step and multi-step reasoning as mentioned in Appendix A.3 and our paper aims at unleashing VoCoT reasoning ability of LMMs without affecting the original abili-

Method	Avg. # Tokens Generated	Avg. # Visual Tokens Added	Avg. Inference Time per Query
VolCano-SE	2.2	0	0.11s
VolCano	115.9	326	0.74s

Table 13: Statistics of single-step and multi-step inference in VSR.

Model	MMBench	SEED	MMMU	MMT	MMVet
LLaVA1.5-7B	64.3	53.8	28.4	45.0	30.5
VolCano-SE	61.1	54.2	29.3	45.1	32.1
VolCano	63.3	57.5	31.2	45.2	32.5

Table 14: Performance of different models with **single-step inference** settings.

ties. To validate that, we evaluate VolCano with the same single-step inference strategy as other compared baselines on general benchmarks in Table 14. It can be seen that VolCano maintains the original ability for conventional single-step inference, performing close to the two baselines in such settings and even surpassing them in some tasks. This indicates that the VoCoT framework does not hurt the original and fundamental capabilities of LMMs.

B.4 Grounding Capabilities of VolCano

Besides the reasoning capability analyzed in Section 4.5, another key capability to ensure the effectiveness of VoCoT is the grounding. However, the commonly adopted RefCOCO (Kazemzadeh et al., 2014) dataset is widely used in training LMMs and not applicable for zero-shot evaluation. Besides, another problem with RefCOCO is that the query is relatively simple. To address this, we consider CLEVR-Ref in the main part because complex queries are considered. As a step further, we are interested in the grounding capability of LMMs during the generation process, but it is difficult to evaluate under this setting. Therefore, we conduct a preliminary exploration.

Specifically, we evaluate the performance of models to produce grounded captions: requiring models to annotate objects with coordinates while describing the images. 100 images are sampled from the LVIS (Gupta et al., 2019) validation set for evaluation. Pairwise evaluation is performed to compare the grounded contents generated by two models. Given the image, ground-truth object information, and responses from 2 models, the judge, GPT-4V, will score 2 responses from multiple perspectives (including both content accuracy and coordinates accuracy) and determine the winner.

Model	LLaVA1.5	LLaVA1.5	VolCano
Visual Input	Image	Image + Object Info.	Image
CLEVR Acc.	43.73	45.70	56.17

Table 15: Performance of models in CLEVR with different visual inputs. Acc. and Info. are short for Accuracy and Information, respectively.

As GPT-4V itself can not generate precise coordinates, we conduct a sanity check whether GPT-4V can evaluate the relevance between two bounding boxes described in texts. For ease of testing, we ask GPT-4V to evaluate responses in RefCOCOg that include single coordinates, and measure the Pearson correlation coefficient between "coordinate accuracy" judged by GPT-4V and the actual IoU. The coefficient is 0.932, indicating that GPT-4V can accurately judge the matching degree between coordinates represented in text. For ease of testing, we ask GPT-4V to evaluate responses in RefCOCOg that include single coordinates, and measure the Pearson correlation coefficient between "coordinate accuracy" judged by GPT-4V and the actual IoU. The coefficient is 0.932, indicating that GPT-4V can accurately judge the matching degree between coordinates represented in text.

Generally, we believe that the current setup can, to some extent, reflect the grounding abilities of models during generation. Ultimately, we compare VolCano with Qwen-VL-Chat and MiniGPTv2. The win rates of VolCano against Qwen-VL-Chat and MiniGPTv2 are 76.5 and 82.0, respectively, indicating that VolCano can perform better in simultaneously locating and describing visual contents.

B.5 Computational Efficiency

VoCoT leads to additional computational overheads compared with traditional single-step reasoning: (1) RefBind only introduces indexing operations without float-point calculations. Additional cost is caused by the visual tokens added to the sequences that will be processed. (2) Multi-step reasoning leads to additional computation by requiring more tokens to be generated. Since precise calculation of the computation cost is challenging, we provide

empirical statistics in Table 13.

Notice that additional computation is inevitable in CoT methods. However, with Flash Attention (Dao et al., 2022) and KV Cache methods used during generation, we found the increase in token quantities does not lead to excessive inference time. In the future, we will follow text CoT papers to explore efficient decoding methods which help improve the efficiency of VoCoT.

B.6 Can Other Open-Source LMMs Directly Utilize Object Information?

Besides the traditional single-step reasoning paradigm, we wonder whether open-source LMMs like LLaVA can utilize the provided ground truth object information to perform grounded reasoning and enhance the improvement. We conduct an experiment on CLEVR where gold object coordinates exist in the dataset. According to Table 15, it is observed that LLaVA benefits from the information but can not utilize it effectively. In contrast, VolCano can perform localization, analysis and reasoning on its own, showing clear superiority.

B.7 Potential Language Bias in Spatial Reasoning

As presented in 3, text-only CoT method performs the best in VSR, we attribute this phenomenon to two reasons: (i) scenarios in VSR are relatively simple, and (ii) the text-only models can better leverage the language bias in spatial relationships as a shortcut. Firstly, each type of object appears only once in an image, with a one-to-one correspondence between the text and the object. so coordinates are not required to resolve ambiguity.

Secondly, as noted in (Kamath et al., 2023), spatial reasoning datasets exhibit some language biases (e.g., a television is more likely to be on a table rather than under it). We find that such biases are more likely to be exploited by text-only CoT models, while VoCoT-based VolCano relies more on analyzed visual information.

We conduct an experiment: replacing the images in VSR with completely black images and using the original queries to ask whether the corresponding spatial relationships exist in the image: (1) The VoCoT-based VolCano predicts "no" for 99% of the samples (in line with expectations). (2) The text-only CoT-based model predicted "yes" for 26% of the samples, achieving a 54.1% accuracy rate

among these predictions (better than random choice for the binary questions). (3) This phenomenon demonstrates that the text-only CoT-based model is more prone to being influenced by language bias, which provides it with a shortcut and additional advantage in VSR.

B.8 Case Study

Different from the black-box single-step reasoning paradigm, VolCano produces interpretable responses with the reasoning paths in text. Please see Figure 8, 9, 10, 11, 12, 13 for cases from representative datasets.

B.9 Discussion on Potential Social Impacts and Bias

We discuss potential issues and our solutions from the following perspectives:

1. **Visual Bias:** object categories in object detection datasets are unevenly distributed (mainly a long-tail distribution). To address the issue, we perform a balanced sampling of images based on the included object categories. For the constructed dataset, we performed manual sampling and inspection and applied some filtering methods mentioned in Appendix A.1 to improve data quality. By checking the final constructed datasets, we did not observe any significant bias issues.
2. **Misinformation:** LMMs may produce erroneous information, namely hallucinations. The design of visually grounded representation in VoCoT aims to mitigate object hallucinations, and the experimental results validate the effectiveness.
3. **Privacy issue:** The images we adopt come from open-source and widely used datasets. Our construction method does not introduce additional privacy risks. Furthermore, we believe our object-centric method can be utilized to detect potential privacy issues in images. In future work, methods like RLHF will be used to guide VolCano to avoid detecting and analyzing data with potential privacy issues.
4. Beyond the above concerns. We utilize existing resources and there is no issue regarding potential personal information leakage and offensive content. The utilized tool, GPT-4V, also possesses the capability to avoid generating offensive content. We manually checked the

constructed dataset to ensure there is no such issues. We will continue to follow current responsible AI methods to monitor and alleviate our model and dataset for any biases or issues.

B.10 Discussion on the Use of Utilized and Presented Artifacts

In this work, we utilize existing artifacts including the data resources (GQA (Hudson and Manning, 2019), COCO (Lin et al., 2014), and LVIS (Gupta et al., 2019)), pre-trained models (CLIP (Radford et al., 2021), Mistral (Jiang et al., 2023), and Qwen2 (Yang et al., 2024)), and existing datasets as listed in Table 1. All utilized artifacts are open-sourced to the research community. We carefully follow the license to use artifacts and ensure they are applicable for the research purpose. All utilized artifacts mainly focus on the English domain while Qwen2 and Mistral both possess multi-lingual capabilities. Please refer to the original resource for other information about the artifacts.

As for the artifacts we presented in this paper, including VoCoT-Instruct-80K and pre-trained VolCano and VolCano_{Q2}, we will release the data, code, and model weights to the community for research purpose. Our introduced artifacts are primarily designed for the English domain and will be extended to more languages. Our artifacts are designed with the principle of universality and fairness, without any preference for specific demographic groups.

B.11 Usage of AI Assistants

In this work, we mainly utilize GPT-4V as the AI assistants for preliminary exploration as in Figure 1, data transforming as in Section 2.2, and as an intelligent agent to judge the performance of models (Appendix B.4). Besides that, we utilize ChatGPT to help polish some parts of this paper.


Opeation	Mapping Rule
relate: sub, relation, obj	Check the {subject} that is {arg2} {object}.
same: attribute [obj1,obj2]	The question ask if the two objects has same {attribute}.
	Check if they have same {attribute}.
common: [obj1,obj2]	The question ask the common attribute of the two objects.
different: attribute, [obj1, obj2]	The question ask if the two objects has different {attribute}
and: [obj1, obj2]	The question ask about 'and' relation.
select: obj1	Find {obj1}
exist: ? obj1	It doesn't exist. if obj1 is not in annotation else It exist
verify: attribute,value, obj1	Verify if the {attribute} of {obj1} is {value}.
or: [obj1, obj2]	The question ask about 'or' relation.
choose: obj1, attribute, value1, value2, obj2	Think {obj1}'s {attribute} is {value1} or {value2} of {obj2}.
choose: obj1, attribute, value1, value2,	Think {obj1}'s {attribute} is {value1} or {value2}.

Table 16: Mapping rule for transferring SQL-like query statement to string in GQA Source Type Data construction.

```

messages = [ {"role": "system", "content": f"""You are an excellent generator of image QA reasoning processes based on question-answer pairs and object information represented by object bounding boxes (x_left_top, y_left_top, x_right_down, y_right_down).Your task is to generate reasoning process based on the questions and answers you are given. The reasoning process should include the reasoning path, relevant object bounding boxes, and inference clues, including but not limited to the object's number, location, and your own background knowledge. The object in your reasoning path must annotate with object bounding box. The bounding box must come from the object information given by the user, please do not detect it yourself! ! ! Don't mention object information directly, just annotate it with bounding boxes.When you refer to the information in prompt, the text should show that you did not know the answer in advance, but that you reasoned it out yourself. And don't directly say that something doesn't appear in the information provided. Don't mention anything in the prompt in your reply, and don't mention bounding boxes in the generated reasoning process.You will follow instructions to the best of your ability. Your response should follow the following format: "Thought":""""}

```



```

{"role": "user", "content": [{"type": "text", "text": "[IMAGE1];[Object Info]: Coffee1: [0.04, 0.25, 0.20, 0.28], Bean: [0.22 , 0.12 , 0.29 , 0.32 ], Vegetable: [0.52, 0.11, 0.29, 0.31 ], Coffee2: [0.76, 0.25, 0.21, 0.27 ], Yam: [0.69, 0.48, 0.27, 0.34 ], Burrito: [0.34, 0.40, 0.35, 0.44 ], Orange: [0.05, 0.50, 0.27, 0.34 ], Chopsticks: [0.61, 0.69, 0.22, 0.31] [question]: What is the food with the most Vitamin C in this image?" [answer]: "Orange."}, {"type": "image_url", "image_url": {"url": "f:data:image/jpeg;base64,{image1}"}}]}

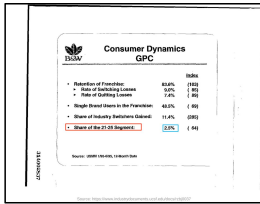
```

```

{"role": "assistant", "content": "" "thought": "From the picture, you can see two cups of coffee [0.04, 0.25, 0.20, 0.28],[0.76, 0.25, 0.21, 0.27 ], one portion of bean[0.22 , 0.12 , 0.29 , 0.32 ], one portion of vegetables[0.34, 0.40, 0.35, 0.44 ], one portion of yam[0.69, 0.48, 0.27, 0.34 ], one burrito[0.34, 0.40, 0.35, 0.44 ], and a plate of oranges[0.05, 0.50, 0.27, 0.34 ]. Among these foods, bean[0.22 , 0.12 , 0.29 , 0.32 ] contains protein, yams[0.05, 0.50, 0.27, 0.34 ] and burrito[0.34, 0.40, 0.35, 0.44 ] are rich in starch, vegetables[0.34, 0.40, 0.35, 0.44 ], and oranges[0.05, 0.50, 0.27, 0.34 ] are foods that may contain vitamin C, but oranges[0.05, 0.50, 0.27, 0.34 ] have a higher vitamin C content, so oranges[0.05, 0.50, 0.27, 0.34 ] are foods that contain more vitamin C."""}

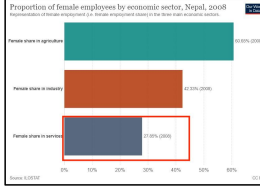
```

Table 17: Prompt and one in-context sample for VQA-Based Source Type Data generation.



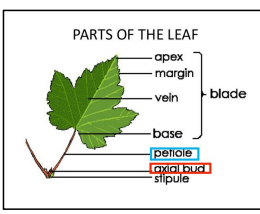
Question: what is the percentage of the share of the 21-25 segment?
Ground Truth: 2.5%
LLaVA : 12
Vocalno-SE : 2.5
Volcano : Find the 21-25 segment [c] 0.23,0.61,0.70,0.70 [c] . Check the percentage of it, got the percentage[c] 0.60,0.60,0.70,0.69 [c] . The percentage[c] 0.60,0.60,0.69,0.69 [c] is 2.5%. So the answer is 2.5%.

Case from DocVQA



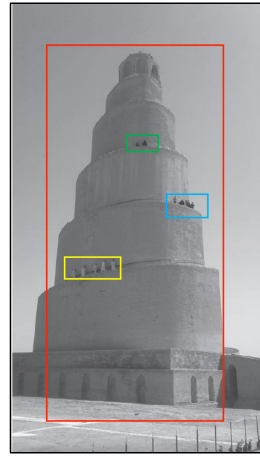
Question: What color represents Female share in service ?
Ground Truth: Grey
LLaVA : Blue
Vocalno-SE : Gray
Volcano : Find the female share in service [c] 0.19,0.60,0.53,0.77 [c] . Check the color representing it, got the color[c] 0.20,0.60,0.53,0.77 [c] . The color[c] 0.20,0.60,0.53,0.77 [c] is gray. So the answer is gray.

Case from ChartQA



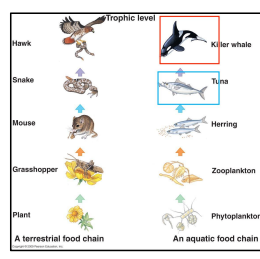
Question: which layer is between axial bud and base? Select from following options: margin; petiole; vein; stipule.
Ground Truth: petiole
LLaVA : margin
Vocalno-SE : petiole
Volcano : Find the axial bud [c] 0.55,0.76,0.75,0.89 [c] . Check the layer between it and base, got layer[c] 0.56,0.62, 0.72,0.75 [c] . The layer is a petiole.[c] 0.56,0.62, 0.72,0.75 [c] So the answer is petiole.

Case from AI2D



Question: The function of this structure was to?
Ground Truth: Signal a time to pray, as in a minaret.
LLaVA : bell tower
Vocalno-SE : bell tower
Volcano : The structure in the image appears to be a large tower [c] 0.22,0.10,0.70,0.92 [c] , which is a common type of structure used for various purposes. Given its size and prominence, it could have served as a landmark or a reference point for navigation. The presence of multiple people [c] 0.45,0.29,0.58,0.35 [c] , [c] 0.61,0.48,0.75,0.48 [c] , [c] 0.24,0.57,0.48,0.62 [c] on the structure suggests that it might have been a point of interest or a gathering place for locals and visitors alike. The structure's design and the way it dominates the image indicate that it could have been a significant architectural feature in the area, possibly representing the cultural or historical heritage of the place.

Case from MMMU



Question: You can see how organisms are interconnected from the diagram given. What will be the effect if all the Killer whales are removed? Select from following options: The population of tuna will increase; Mouse will decrease in number; The phytoplankton will decrease; The grasshopper will die.
Ground Truth: The population of tuna will increase
LLaVA : The phytoplankton will decrease
Vocalno-SE : The phytoplankton will decrease
Volcano : To determine the effect of removing all Killer whales, we look at the trophic level diagram. The Killer whale [c] 0.60,0.05,0.81,0.21 [c] is at the top of the food chain, indicating it is apex predators. The removal of these apex predators would have cascading effects down the food chain. The removal of Killer whales would disrupt the balance of the ecosystem. As apex predators, they help control the populations of their prey, such as the Tuna [c] 0.58,0.29,0.79,0.35 [c] , among others. The absence of these predators would lead to overpopulation of prey species.

Case from MathVista

Figure 7: Cases on additional benchmarks included in Appendix B.2.

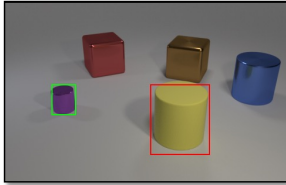
```
messages = [ {"role": "system", "content": f"""You are an excellent image describer and question-answer generator based on the image and object information which is represented by object bounding box (x_left_top, y_left_top, x_right_down, y_right_down). You have three tasks in total. Your first task is to ask a complex question that requires close inspection of the image and strong reasoning ability to answer. Your second task is to answer the question you raised solely based on the given image. Your third task is to generate the reasoning thought. The reasoning thought should contain the reasoning path, relative object bounding box and inference clue, include but are not limited to the object numbers, location and background knowledge from yourself. The bounding box must come from the user given object information, Do not detect by yourself !!!! Do not mention the object information directly, just annotate with bounding box. When you ask questions, try to find the most valuable information in the picture to ask about, and ask a question that is relevant to that information. When you ask questions, do not involve violence, advertisement, possible invasion of privacy, or questions that may cause discomfort. Do not mention anything from the prompt in your response and Do not mention bounding box in your generated question. You will follow the instructions to the best of your ability. Your response should follow the following format: "question":"","answer":"","Thought":""""},
```



```
 {"role": "user", "content": [{"type": "text", "text": "[IMAGE1]:[Object Info]: Coffee1: [0.04, 0.25, 0.20, 0.28], Bean: [0.22, 0.12, 0.29, 0.32], Vegetable: [0.52, 0.11, 0.29, 0.31], Coffee2: [0.76, 0.25, 0.21, 0.27], Yam: [0.69, 0.48, 0.27, 0.34], Burrito: [0.34, 0.40, 0.35, 0.44], Orange: [0.05, 0.50, 0.27, 0.34], Chopsticks: [0.61, 0.69, 0.22, 0.31]"}, {"type": "image_url", "image_url": {"url": "data:image/jpeg;base64,{image1}"}}]},
```

```
 {"role": "assistant", "content": "" "question": "What is the food with the most Vitamin C in this image?", "answer": "Orange.", "thought": "From the picture, you can see two cups of coffee [0.04, 0.25, 0.20, 0.28],[0.76, 0.25, 0.21, 0.27], one portion of bean[0.22, 0.12, 0.29, 0.32], one portion of vegetables[0.34, 0.40, 0.35, 0.44], one portion of yam[0.69, 0.48, 0.27, 0.34], one burrito[0.34, 0.40, 0.35, 0.44], and a plate of oranges[0.05, 0.50, 0.27, 0.34]. Among these foods, bean[0.22, 0.12, 0.29, 0.32] contains protein, yams[0.05, 0.50, 0.27, 0.34] and burrito[0.34, 0.40, 0.35, 0.44] are rich in starch, vegetables[0.34, 0.40, 0.35, 0.44], and oranges[0.05, 0.50, 0.27, 0.34] are foods that may contain vitamin C, but oranges[0.05, 0.50, 0.27, 0.34] have a higher vitamin C content, so oranges[0.05, 0.50, 0.27, 0.34] are foods that contain more vitamin C.""}],
```

Table 18: Prompt and one in-context sample for Image-Only Source Type Data generation.

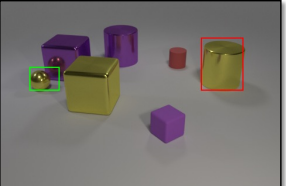
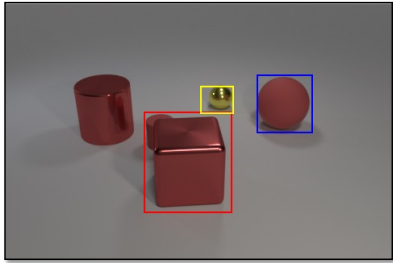


Question: The small thing to the left of large yellow object has what shape?

Volcano: Prediction with VoCoT : Find the large yellow cylinder [c] 0.51,0.45,0.72,0.70 [c]. Check the small cylinder to the left of it, got the cylinder [c] 0.16,0.45,0.25,0.56 [c] . The cylinder is round. So the answer is round.

Question: There is a sphere to the left of the red matte object to the right of the red block; what is its color?

Volcano: Find the red block [c] 0.360,0.446,0.586,0.708 [c]. Check the red matte object to the right of it, got the object [c] 0.65,0.35,0.79,0.50 [c] . Check the sphere to the left of it, got the sphere [c] 0.51,0.38,0.59,0.45 [c] . The sphere [c] 0.51,0.38,0.59,0.45 [c] is gold. So the answer is gold.



Question: There is a sphere that is the same material as the yellow cylinder; what is its size?

Volcano: Find the yellow cylinder [c] 0.716,0.306,0.866,0.492 [c] . Check the sphere has the same material, got the sphere [c] 0.106,0.410,0.212,0.492 [c] . The sphere is small. So the answer is small.

Question: The purple object behind the rubber sphere in front of the small green cylinder is made of what material?

Volcano: Find the small green cylinder [c] 0.57,0.35,0.64,0.44 [c] . Check the rubber sphere in front of it, got the sphere [c] 0.35,0.50,0.45,0.60 [c] . Check the purple cylinder behind it, got the cylinder [c] 0.44,0.38,0.52,0.48 [c] . The cylinder is made of plastic. So the answer is plastic.

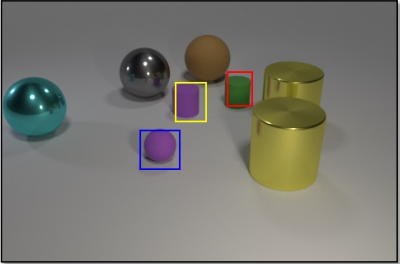
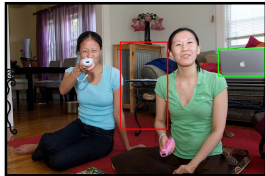


Figure 8: Cases on CLEVR dataset.



Question: Does the computer to the right of the shelf look black or silver?

Volcano : Find the shelf [c] 0.43,0.31,0.62,0.65 [/c] . check the computer to the right of it, got the computer [c] 0.81,0.33,1.00,0.44 [/c]. think the computer [c] 0.81,0.33,1.00,0.44 [/c] is silver or black. the answer is silver.

Question: On which side is the small nut?

Volcano : Find the small nut [c] 0.78,0.59,0.90,0.64 [/c] . the nut [c] 0.78,0.59,0.90,0.64 [/c] is on the right of the image. so the answer is right.

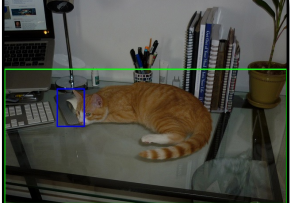


Question: What animal is standing in front of the giraffes?

Volcano : Find the giraffes [c] 0.35,0.38,0.62,0.53 [/c] . check the animal standing in front of it, got the animal [c] 0.68,0.51,0.86,0.62 [/c] . the zebra [c] 0.68,0.51,0.86,0.62 [/c] is standing in front of the giraffes. [c] 0.35,0.38,0.62,0.53 [/c] so the answer is zebra.

Question: Is the mouse on top of a desk?

Volcano : Find the computer mouse [c] 0.18, 0.45, 0.28, 0.59 [/c] . verify if the computer mouse [c] 0.18, 0.45, 0.28, 0.59 [/c] is on top of the desk [c] 0.00, 0.38, 1.00, 0.87 [/c] . the answer is yes.



Question: Does the pipe have a different color than the fridge?

Volcano : find the pipe [c] 0.05, 0.20, 0.16, 0.26 [/c] . find the refrigerator [c] 0.44,0.47,0.54,0.64 [/c] . the question ask if the two objects has different color. both the pipe [c] 0.05, 0.20, 0.16, 0.26 [/c] and the fridge [c] 0.44, 0.47, 0.54, 0.64 [/c] are white. so the answer is no.

Figure 9: Cases on GQA dataset.



Question: What object is depicted in the image?

Option:

- A. A painting
- B. A fish tank
- C. A birdhouse
- D. A statue of an angel holding a book

Ground Truth: A statue of an angel holding a book

Volcano: Find the object [c] 0.31,0.18,0.64,0.93 [c]. The object [c] 0.31,0.18,0.64,0.93 [c] is a statue.[c] 0.31,0.18,0.64,0.93 [c] So the answer is statue.

Question: Why are the photographers taking pictures of the woman?

Option:

- A. She is conducting an outdoor class
- B. They are documenting a car accident
- C. She is famous
- D. She is a person of interest in an investigation

Ground Truth: She is famous



Volcano: The photographers are taking pictures of the woman because she is a celebrity, as indicated by her prominent position in the image and the attention she is receiving. The presence of multiple photographers [c] 0.26, 0.10, 0.42, 0.29 [c], [c] 0.53, 0.14, 0.66, 0.30 [c], [c] 0.77, 0.03, 0.87, 0.43 [c], [c] 0.13, 0.10, 0.23, 0.30 [c], [c] 0.64, 0.14, 0.77, 0.29 [c] around her suggests that she is the focal point of their attention. This scenario is typical in celebrity events or appearances where fans and media gather to capture moments with the person of interest. The fact that she is the center of attention, surrounded by photographers, implies that she is a public figure, likely known for her fame or achievements, which is why she is being photographed



Question: What is the color of the tablecloth on the table?

Option:

- A. Blue
- B. Brown
- C. Red
- D. White

Ground Truth: White

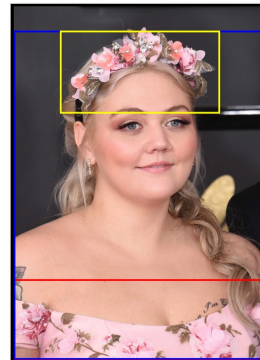
Volcano: Find the table [c] 0.22, 0.51, 0.82, 0.83 [c]. Check the tablecloth on it, got the tablecloth [c] 0.22,0.51, 0.82, 0.83 [c]. The tablecloth is white. So the answer is white.

Question: What color is the flower crown that the woman in the pink dress is wearing?

Option:

- A. Red
- B. White
- C. Yellow
- D. Pink

Ground Truth: Pink



Volcano: Find the pink dress [c] 0.14,0.77,0.85,1.00 [c]. Check the woman in it, got the woman [c] 0.14, 0.06, 0.85, 0.99 [c]. Check the flower crown that it wearing, got the flower crown [c] 0.2750.06,0.72,0.30 [c]. The flower crown is pink. So the answer is pink.

Figure 10: Cases on Seed-Bench dataset. Seed-Bench provide a question and a list of answer options for each sample.



Question: What is the position of the wooden table in the room?

Option:

- A. Against the wall
- B. Next to the couch
- C. In the center
- D. In the corner

Ground Truth: In the center

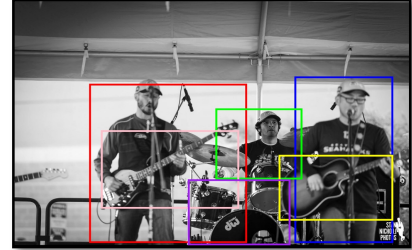
Volcano: Find the room [c] 0.00, 0.12, 1.00, 0.87 [c]. Check the wooden table in it, got the table [c] 0.24, 0.64, 0.87, 0.87 [c]. The table [c] 0.24, 0.64, 0.87, 0.87 [c] is in the middle. So the answer is center.

Question: What is the overall theme of this image?

Option:

- A. An outdoor camping trip
- B. A musical performance
- C. A motorcycle event
- D. A street market

Ground Truth: A musical performance



Volcano: The overall theme of this image is centered around a musical performance, specifically a concert. This conclusion is drawn from the presence of multiple individuals [c] 0.19, 0.39, 0.59, 0.82 [c], [c] 0.71, 0.37, 0.96, 0.82 [c], [c] 0.51, 0.46, 0.73, 0.65 [c] who are positioned in a manner that suggests they are performing together, likely on stage. The presence of musical instruments, such as guitars [c] 0.67, 0.58, 0.96, 0.76 [c], [c] 0.22, 0.52, 0.57, 0.73 [c], and a drum set [c] 0.44, 0.65, 0.70, 0.83 [c] further supports this theme. The arrangement and focus of these elements, along with the attire of the individuals, which is not explicitly detailed but can be inferred from the context of a concert, suggest a live performance. The setting, likely a stage, and the engagement of the individuals with their instruments, indicate a musical event where the performers are actively playing for an audience. This setup is typical of a concert, where musicians come together to showcase their talents and entertain their listeners.



Question: What is the dominant color on the upper part of the image?

Option:

- A. Blue
- B. Brown
- C. Red
- D. White

Ground Truth: White

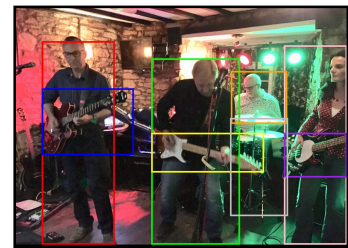
Volcano: Find the upper part of the image [c] 0.00, 0.21, 0.99, 0.39 [c]. The dominant color on the upper part of the image [c] 0.00, 0.21, 0.99, 0.39 [c] is blue. So the answer is blue.

Question: How many people are playing musical instruments in the image?

Option:

- A. 1
- B. 4
- C. 3
- D. 2

Ground Truth: 4



Volcano: In the image, there are four individuals playing musical instruments. The first person [c] 0.08, 0.26, 0.30, 0.87 [c] is playing a guitar [c] 0.08, 0.40, 0.35, 0.60 [c]. The second person [c] 0.41, 0.31, 0.67, 0.87 [c] is also playing a guitar [c] 0.41, 0.53, 0.75, 0.65 [c]. The third person [c] 0.81, 0.27, 1.00, 0.87 [c] is playing a bass guitar [c] 0.81, 0.53, 1.00, 0.67 [c]. The fourth person [c] 0.65, 0.35, 0.82, 0.51 [c] is playing a drum set [c] 0.64, 0.49, 0.82, 0.78 [c].

Figure 11: More Cases on Seed-Bench dataset.



Statement: The cat is in the backpack.
Ground Truth: no

Volcano: The image shows a cat [c] 0.16, 0.18, 0.83, 0.83 [/c] lying on a bench [c] 0.00, 0.12, 0.90, 0.87 [/c]. However, there is no any indication that the cat is in the backpack. Therefore, the answer is no.

Statement: The cat is surrounding the remote.
Ground Truth: no

Volcano: In the image, the cat [c] 0.18, 0.12, 1.00, 0.87 [/c] is lying next to the remote [c] 0.002, 0.484, 0.696, 0.874 [/c], but it is not surrounding the remote. Instead, the cat is lying beside the remote. Therefore, the answer is no.

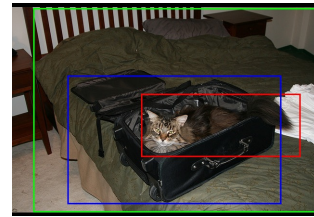
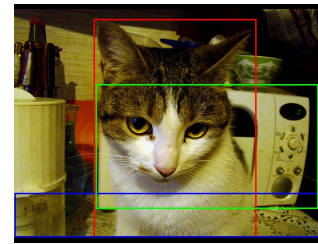


Statement: The suitcase contains the book.
Ground Truth: yes

Volcano: The image shows a suitcase [c] 0.00, 0.57, 0.78, 0.81 [/c] with a book [c] 0.00, 0.16, 0.63, 0.61 [/c] inside it. The book is placed inside the suitcase, which implies that the event "the suitcase contains the book" is taking place in the image. Therefore, the answer is yes.

Statement: The cat is in the microwave.
Ground Truth: no

Volcano: The image shows a cat [c] 0.26, 0.15, 0.79, 0.87 [/c] sitting on a counter [c] 0.00, 0.73, 1.00, 0.85 [/c] next to a microwave [c] 0.26, 0.37, 0.99, 0.78 [/c]. However, there is no evidence of the cat being inside the microwave. Therefore, the answer is no.



Statement: The cat is within the suitcase.
Ground Truth: yes

Volcano: In the image, the cat [c] 0.42, 0.44, 0.93, 0.65 [/c] is indeed inside the open suitcase [c] 0.18, 0.38, 0.87, 0.81 [/c]. The cat is sitting inside the suitcase, which is placed on the bed [c] 0.07, 0.17, 1.00, 0.83 [/c]. Therefore, the event "the cat is within the suitcase" is taking place in the image. The answer is yes.

Statement: The book is facing the cat.
Ground Truth: yes

Volcano: In the image, the book [c] 0.16, 0.42, 0.99, 0.83 [/c] is open and facing the cat [c] 0.09, 0.19, 0.63, 0.63 [/c]. The cat is laying on the book, and the book is positioned in such a way that the cat is facing it. Therefore, the event "the book is facing the cat" is taking place in the image. The answer is yes.

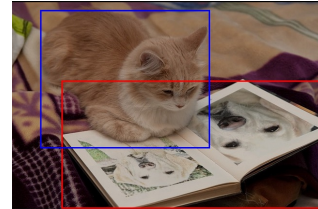
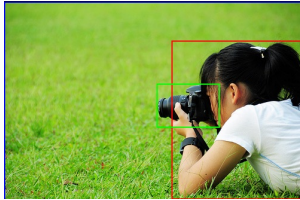


Figure 12: Cases on VSR dataset. VSR dataset provides a statement for each image. The task is to judge whether the statement is right about this image.

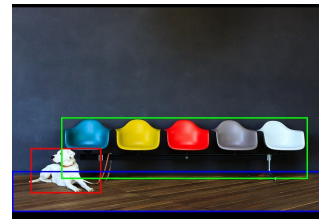


Volcano-SE : The image features a young woman lying on her stomach in a grassy field, taking a photo with her camera. She is **positioned in the center of the scene**, with her camera held up to her face, capturing the beauty of the green grass around her. The camera is clearly visible in her hands, as she focuses on her subject. The woman appears to be enjoying the outdoor setting and the opportunity to capture a memorable moment.

Volcano : The image features a young woman [c] 0.56, 0.29, 1.00, 0.82 [c] lying on the grass [c] 0.00, 0.16, 1.00, 0.83 [c] and taking a picture. She is holding a camera [c] 0.51, 0.44, 0.72, 0.58 [c] in her hands, focusing on capturing the moment. The woman's position and the camera's position suggest she is actively engaged in taking .

Volcano-SE : The image features a white dog sitting on a wooden floor in front of a row of colorful chairs. There are five chairs in total, each with a different color, including blue, yellow, red, and **two gray chairs**. The chairs are arranged in a line, creating a visually appealing display. The dog appears to be calmly sitting and enjoying its surroundings.

Volcano : The image features a white dog [c] 0.06, 0.62, 0.29, 0.77 [c] sitting on the floor [c] 0.00, 0.70, 1.00, 0.83 [c] in front of a row of colorful chairs [c] 0.16, 0.52, 0.96, 0.72 [c] . The chairs are arranged in a line, with each chair having a different color. The dog is positioned between the chairs, and it appears to be calmly sitting in the space.



Volcano-SE : The image features a black dog sitting on a sandy beach, enjoying the sunny day. The dog appears to be a medium-sized breed, possibly a Labrador or a similar type. It is wearing a blue collar, which adds a pop of color to the scene. The dog is sitting on the sand, with its front **paws positioned close to each other**, and it seems to be looking up at the camera with a happy expression.

Volcano : The image features a black dog [c] 0.25, 0.21, 0.70, 0.76 [c] sitting on the sand [c] 0.00, 0.22, 1.00, 0.78 [c] . The dog appears to be smiling, giving the impression that it is happy and enjoying its time on the beach. The dog's position and expression convey a sense of contentment and relaxation.



Volcano-SE : The image features a dirt road stretching through a rural area, with a few **scattered trees visible in the background**. The road appears to be empty, with no visible vehicles or people. The scene is quite peaceful, with the dirt road leading into the distance and the trees providing a natural backdrop.

Volcano : The image features a dirt road [c] 0.00, 0.07, 1.00, 0.92 [c] with a clear view of the ground. The road appears to be empty and devoid of any vehicles or people. The dirt road is surrounded by a natural landscape, giving the impression of a rural or remote area. The road's surface is uneven and consists of dirt, which is typical for such environments.



Figure 13: Cases from the AMBER dataset. The task is to describe each image. We both present our VolCano and VolCano-SE responses for these cases. VolCano-SE is trained without VoCoT data. The underline phrase is hallucination generated by VolCano-SE.