Compositional Chain-of-Thought Prompting for Large Multimodal Models

Chancharik Mitra Brandon Huang Trevor Darrell Roei Herzig
University of California, Berkeley

Abstract

The combination of strong visual backbones and Large Language Model (LLM) reasoning has led to Large Multimodal Models (LMMs) becoming the current standard for a wide range of vision and language (VL) tasks. However, recent research has shown that even the most advanced LMMs still struggle to capture aspects of compositional visual reasoning, such as attributes and relationships between objects. One solution is to utilize scene graphs (SGs)—a formalization of objects and their relations and attributes that has been extensively used as a bridge between the visual and textual domains. Yet, scene graph data requires scene graph annotations, which are expensive to collect and thus not easily scalable. Moreover, finetuning an LMM based on SG data can lead to catastrophic forgetting of the pretraining objective. To overcome this, inspired by chain-of-thought methods, we propose Compositional Chain-of-Thought (CCoT), a novel zero-shot Chain-of-Thought prompting method that utilizes SG representations in order to extract compositional knowledge from an LMM. Specifically, we first generate an SG using the LMM, and then use that SG in the prompt to produce a response. Through extensive experiments, we find that the proposed CCoT approach not only improves LMM performance on several vision and language (VL) compositional benchmarks but also improves the performance of several popular LMMs on general multimodal benchmarks, without the need for fine-tuning or annotated ground-truth SGs. Code: https://github.com/chancharikmitra/CCoT.

1. Introduction

In recent years, *Large Multimodal Models* (LMMs) such as LLaVA [46], GPT-4V [55], and InstructBLIP [16] have demonstrated impressive results in the field of vision and language (VL), especially in multimodal reasoning and visual question-answering (VQA) [5, 39, 47, 48, 52]. However, recent empirical studies [18, 28, 51] show that the best-performing VL models tend to view images as a "bag of objects". Consider the following example in Figure 1. Suppose a VL model is asked to describe the provided image. The provided image contains many objects: a laptop, a mouse, some books, and a table. It is a challenging question to describe exactly how these objects are situated in relation to one another as well as their notable characteristics. Thus, we are motivated to utilize the SG, which captures the objects' important relationships and attributes. For example, the LMM uses the generated SG to produce the description: "On a table, there is a stack of books on a laptop."

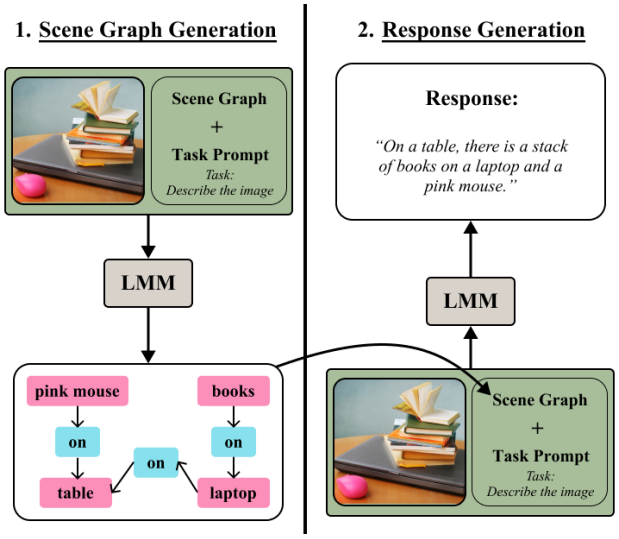

Figure 1. **A high-level overview of our Compositional Chain-of-Thought (CCoT) approach.** Our CCoT method consists of a two-step prompting process: (1) First, the LMM is prompted to generate a scene graph relevant to the image and task prompt, such as the task in the figure "Describe the image". (2) Following this, the LMM is prompted with the generated scene graph, the image, and the task prompt as context for responding in a way that incorporates the compositional information in the scene graph to provide a correct description of the complex scene.

Comprehending the structure of visual scenes is a core issue in machine perception. Visual scenes consist not only of objects but also include relevant characteristics and relationships that are significant to understanding the scenes' compositionality better. In this paper, we consider how to best improve the compositionality of LMMs. Recently, scene graph (SG) annotations—structured graph representations of visual scenes–have been introduced as powerful

VL representations, and have been extensively explored in many previous works [24, 34, 79, 80]. However, SG data is less readily available than textual descriptions as obtaining SGs is costly and thus not scalable.[1] Moreover, training on SG data can lead to forgetting on the pretrained objectives as shown in [28]. Therefore, in this paper, we propose leveraging scene graph representations for LMMs *without annotated scene graph data* and *without finetuning*.

Recently, Large Language Models (LLMs) showed promising results by incorporating Chain-of-Thought (CoT) prompting methods [36, 76]. CoT methods use an LLM to perform a task with intermediate reasoning steps, either zero-shot—with no explicit examples—or few-shot—with explicit examples. Inspired by this, we design a zero-shot, CoT method that utilizes scene graph representations for multimodal and compositional visual reasoning tasks. Our approach allows us to extract more *compositional* knowledge out of an LMM compared to without prompting. Next, we ask ourselves how should we design a CoT prompt method that utilizes the scene graphs without relying on ground truth SG annotations or model finetuning.

Our proposed designed approach—Compositional Chain-of-Thought (CCoT)—can be broken into two steps. The first step is to generate a scene graph in order to circumvent the need for ground truth SG data by using the input image and task prompt (e.g., visual question). The second step is to prompt the LMM with the image, task prompt, and the generated scene graph to produce a response. Incorporating the scene graph in the prompt eliminates the need for fine-tuning and prevents forgetting. Another benefit of our method is that generated SGs can describe any visual scene, therefore making CCoT generally applicable to a wider range of VL tasks. Finally, the fact that the generated scene graphs are compact linguistic representations of images makes CCoT a token-efficient prompting method. This is significant given the limited textual context lengths that LMMs often face due to processing both image and text inputs.

To summarize, our main contributions are as follows: (i) We introduce CCoT, a zero-shot Chain-of-Thought approach that utilizes scene graph representations in order to extract *compositional* knowledge out of an LMM; (ii) Our proposed CCoT method was designed without the need for task-specific fine-tuning or annotated SG data, as well as being applicable and easy to use on various different LMM architectures; (iii) Our method shows improved performance for LLaVA-1.5, Instruct-BLIP, SPHINX, and GPT-4V not only on VL compositional benchmarks like Winoground and WHOOPS! but also on general multimodal benchmarks like SEEDBench, MMBench, and LLaVA-Bench-in-the-Wild highlighting the effectiveness of our approach.

[1]For example, Visual Genome [37] contains only $\sim 100K$ image-SG pairs, which is smaller than the existing LMMs pretraining datasets.

2. Related Work

Large Multimodal Models (LMMs). The development of LMMs is largely the result of pairing LLMs' powerful reasoning capabilities [15, 60, 67] with existing VL models. A good example of such models is contrastive vision and language models [20, 40, 59], which have been a significant step forward in connecting vision and language representations. However, these methods are limited in their direct application to downstream tasks that require a generative component or more explicit reasoning over both modalities, e.g., visual question-answering [5, 23, 29, 31, 52, 61]. The solution came in the form of applying the reasoning and generative capabilities of LLMs to both textual *and* visual information—resulting in the development of LMMs.

LMMs directly reason over embedded visual features [1, 7, 16, 19, 21, 41, 45, 46, 83, 84, 92]. Particularly crucial for the success of these methods is visual instruction finetuning of the model [46, 89]. Inspired by text-only instruction tuning of LLMs [75], visual instruction tuning has been shown effective for complex visual tasks by passing detailed text descriptions and object location information to top-of-the-line LLMs (e.g. GPT-4 [55]). However, this approach requires high-quality training data, which is not always available or scalable. In this paper, we present an approach that eliminates the need for training data.

Similar to LMMs, another class of multimodal methods use code generation as a proxy for visual reasoning (e.g., ViperGPT [65], VisProg [22], and CodeVQA [64]), which we refer to in this paper as *Visual Programmatic Models (VPMs)* [49, 57, 62, 63, 77]. Inspired by Neural Modular Network architectures [3, 4, 33] that leverage and scale the compositional nature of visual reasoning, VPMs build on the recent advent of highly capable out-of-the-box LLMs without the need for additional programming. Notably, these methods do not directly reason over the visual information and are limited by the exact APIs or models they are provided access to via their limited context. Unlike these methods, here we explored the potential of LMMs, which utilize scene graphs as a bridge between the visual and language domains for compositional visual reasoning.

Multimodal Prompting Methods. Considering the growing popularity of LLMs and LMMs, prompting methods have been critical to harnessing their power as they enable precise control over model outputs and provide context within which models can be used. More importantly prompting methods occur at inference time. They include zero-shot methods [35, 69, 71], few-shot methods [13, 17, 50, 54], expert prompting [78], and Chain-of-Thought (CoT) [76, 87], with extensions like self-consistency [73], Tree-of-Thought (ToT) [81], and Graph-of-Thought (GoT) [11, 38, 82] for more complex structures.

To the best of our knowledge, three methods—VidIL [74], DDCoT [91], and Multimodal-CoT ap-

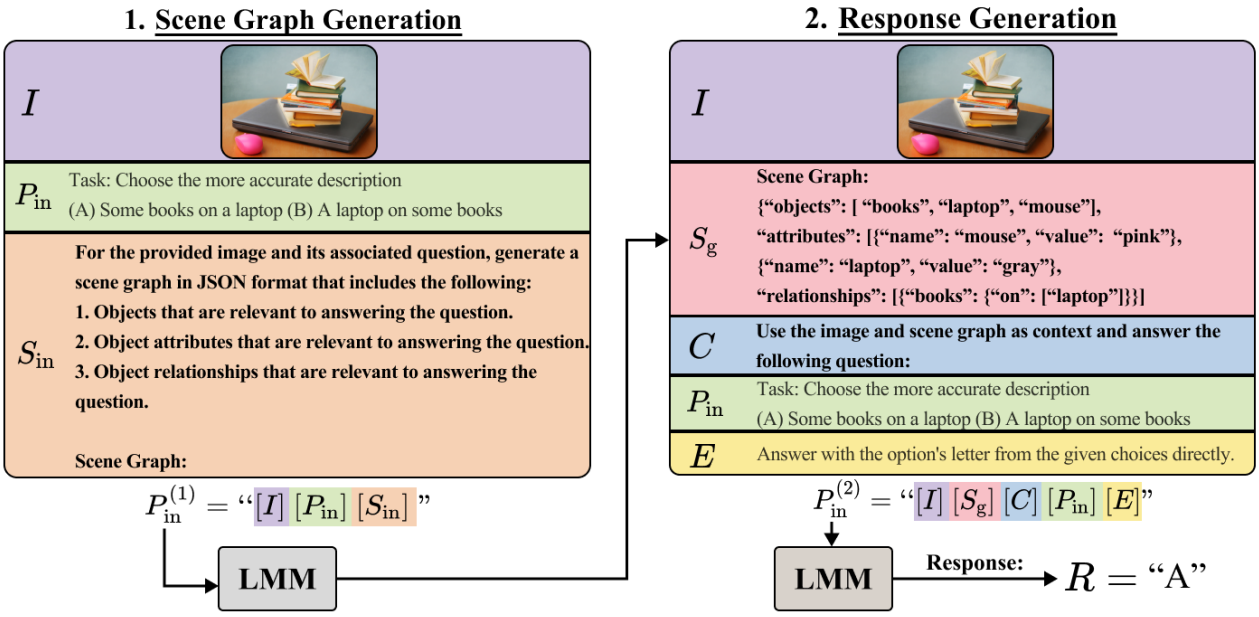

Figure 2. **Full prompt example of CCoT.** The first step in our prompting method is to generate a scene graph given both the image *and* textual task as context. Following this, the answer is extracted by prompting the LMM with the image, scene graph, question, and answer extraction prompt. Prompt sections unique to our method are **bolded**.

proaches [70, 88]—represent the current state-of-the-art in multimodal prompting. VidIL, an architecture specifically designed for video has a language model, which reasons over captions of video frames. Similary, DDCoT designs its own CoT prompting method over image captions rather than explicit visual features. Finally, while Multimodal-CoT leverages an LMM that reasons directly over visual and text input features, its Chain-of-Thought prompting method requires finetuning on ground truth natural language reasoning, which is both annotation and computation costly.

A key difference between CCoT and these methods is that we utilize generated SG instead of captions (generated or collected ground-truth) as a reasoning step in our CoT design. This improves the compositionality of LMMs, which explicitly reason over visual features as well. Additionally, we demonstrate that our method enhances multimodal reasoning more broadly as well. Last, as CCoT is a zero-shot method used at inference time, it is broadly applicable to a wide range of LMM-based architectures.

Compositionality. Compositionality, or the understanding of concepts as being composed of their respective subparts and relationships, is a valuable paradigm for visual concepts via reasoning over the objects, relationships, and attributes in an image. Compositionality has been applied in a variety of domains including: vision and language[2, 14, 18, 28, 42, 66, 85], visual question answering [29, 37, 52], video understanding[6, 8, 25, 27, 53, 72], relational reasoning [9, 10, 30], and scene graphs[24, 26, 32, 58, 79]. Recent empirical studies [28, 68, 86, 90], have shown have that even the strongest LMMs struggle to perform compositional visual understanding, including identifying object attributes and inter-object relations. Specifically, it has been shown that VL models [51] tend to learn a "bag of objects" representation, leading them to be less compositional. In this work, we show that a more structured CoT approach leads to improved compositional reasoning in LMMs, evidenced by improved performance on compositional benchmarks.

3. Compositional Chain-of-Thought

To address the challenge of LMMs viewing images as a "bag of objects," as shown in previous works, our method introduces a novel approach to enhance compositional visual understanding. We begin by describing the standard LMM architecture (Section 3.1). We then introduce our two-step chain-of-thought approach: first is scene graph generation (Section 3.2) and second is response generation (Section 3.3). Our method is illustrated in Figure 2.

3.1. Preliminaries

LMMs are multimodal models that directly reason over both vision and language modalities. They are typically given inputs of one image I and an associated task prompt in text form P_{in} (e.g., questions, caption generation, etc.). Each modality is then encoded into a shared embedding space that a language model $f_\theta(\cdot)$ (parameterized by θ) can reason over. More concretely, the image is encoded using a trainable vision encoder $v_\phi(\cdot)$ (parameterized by ϕ), while

the task prompt is tokenized and then encoded using a fixed language embedding l. Given an input image I and input task prompt P_{in}, the language model (typically an LLM) then outputs a text response R.

$$R = f_\theta(v_\phi(I), l(P_{\text{in}})) \tag{1}$$

The exact LMM sub-modules of the LLM, vision encoding architecture, and pretraining method for parameters θ, ϕ differ between models but the overarching method described above remains the same.

We propose CCoT, a zero-shot chain-of-thought prompting method that leverages scene graph generation to improve an LMM's compositional visual understanding and multimodal reasoning. Notably, this method does not require any finetuning as it is purely prompting-based. Furthermore, no annotated SGs are required as the method is zero-shot. Ultimately, our method is centered around a scene-graph generation prompt S_{in} that can be integrated into P_{in} such that the LMM can output a scene graph S_{g} as an intermediate multimodal reasoning step to output better responses to the task prompts, such as questions, classification, or caption generation.

3.2. Step 1: Scene Graph Generation

Our first step is to generate a scene graph S_{g}, obviating the need for ground truth annotated SG data. The scene graph generation prompt S_{in} instructs the LMM to systematically construct a scene graph with three key properties: the *objects*, their *attributes*, and the *relationships* between them. To address the "bag-of-objects" problem, we would like to have a global view of not just the objects, which are the primary units for visual reasoning, but also their properties and how they interact with one another.

In the scene graph generation prompt S_{in}, we further condition its format to be in JSON. This standardization in JSON format is intended to facilitate easier interpretation by the LMM. By systematically organizing visual information through the inclusion of objects, relationships, and attributes in the scene graphs, we enable more structured and comprehensive reasoning. The full prompt, showcasing this structured approach, is illustrated in Figure 2. The scene graph generation method represents a core novel contribution of our work, aiming to overcome the limitations of existing multimodal reasoning models and enhance the compositional understanding of LMMs.

We include both the image I and task prompt P_{in} along with S_{in} to condition the generated scene graph to be relevant to the given task prompt. This is because SGs are inherently very long-tailed: a generated scene graph that is conditioned only on the image, might incorporate information unrelated to the given task prompt.

The entire first prompt to the LMM, which we denote as $P_{\text{in}}^{(1)}$ is constructed by combining the input image I, task prompt P_{in}, and most notably the scene-graph generation prompt S_{in} (showed in red under Scene-Graph Generation in Figure 2). The full prompt is as follows:

$$P_{\text{in}}^{(1)} = \text{“}[I]\ [P_{\text{in}}]\ [S_{\text{in}}]\text{”} \tag{2}$$

where $[\cdot]$ indicates slots for inserting the individual elements of the prompt. The LMM thus generates a SG as follows:

$$S_{\text{g}} = f(v_\phi(I), l(P_{\text{in}}^{(1)})) \tag{3}$$

3.3. Step 2: Response Generation

To bypass the need for finetuning and thus eliminate forgetting, we utilize the generated scene graph S_{g} as an intermediate chain-of-thought reasoning step. The LMM is thus prompted with the original task prompt, image, and corresponding generated scene graph so that all three can be jointly used as context to respond to this new task prompt. The overall input prompt for response generation is thus given as follows:

$$P_{\text{in}}^{(2)} = \text{“}[I]\ [S_{\text{g}}]\ [C]\ [P_{\text{in}}]\ [E]\text{”} \tag{4}$$

In addition to the input image I, original task prompt P_{in}, and generated scene graph S_{g}, we insert a context sentence C and an answer extraction sentence E. C briefly instructs the LMM to use the provided context. Concretely, this is given by "Use the image and scene graph as context and answer the following question:". Finally, while the flexibility of LLM text generation is a great modeling choice for high-level multimodal reasoning, this flexibility also makes response generation in a specific format non-trivial. Many multimodal benchmarks are in a multiple-choice format, for example. Since we evaluate our method on these types of benchmarks, a short additional sub-prompt E (usually a conditioning sentence) is required to return the answer as a letter. For example, our answer extraction sub-prompt "Answer with the option's letter from the given choices directly" is taken from LLaVA-1.5 [45] as it has been shown to be reliable on large multiple-choice benchmarks. However, this method can be easily generalized to other answer formats like short answers or detailed descriptions by modifying or completely removing E. Thus, the LMM generates a final response R to the original image, task prompt pair (I, P_{in}) as follows:

$$R = f(v_\phi(I), l(P_{\text{in}}^{(2)})) \tag{5}$$

4. Experiments and Results

We apply our CCoT approach to four popular LMMs: InstructBLIP-13B [16], LLaVA-1.5-13B [45], Sphinx [44], and GPT-4V [55]. We also evaluated our approach to several baselines across different benchmarks, focusing on multimodal reasoning and VL compositional tasks. Additional results can be found in our Supplementary Section A.

Model	Multimodal Benchmarks			VL Compositional Benchmarks			
	SEED-I	MMBench	LLaVA-W	Wino-Text	Wino-Image	Wino-Group	WHOOPS! VQA BEM
CLIP	-	-	-	30.7	10.5	8.0	-
BLIP	-	-	-	39.0	19.2	15.0	39.0
BLIP2	46.4	-	-	42.0	23.8	19.0	55.0
SGVL†	-	-	-	42.8†	28.5†	23.3†	-
mPlug-OWL2	57.8	64.5	-	-	-	-	-
QwenVL-Chat	58.2	61.2	-	-	-	-	-
InstructBLIP-13B	48.2	36.0	47.2	12.8	13.3	4.5	48.3
InstructBLIP-13B-ZS-CoT	37.6	25.3	45.4	15.8	14.8	6.0	43.36
InstructBLIP-13B-CCoT	56.9 (+8.7)	40.3 (+4.3)	47.9 (+0.7)	26.0 (+13.2)	27.0 (+13.7)	11.5 (+7.0)	62.9 (+14.6)
LLaVA-1.5-13B	68.2	67.0	73.5	33.5	35.0	17.3	47.3
LLaVA-1.5-13B-ZS-CoT	66.7	66.0	68.5	36.8	35.0	19.8	46.6
LLaVA-1.5-13B-CCoT	69.7 (+1.5)	70.7 (+3.7)	74.9 (+1.4)	39.8 (+6.3)	37.3 (+2.3)	22.3 (+5.0)	61.2 (+13.9)
Sphinx	71.6	65.9	70.0	29.0	29.0	16.3	50.0
Sphinx-ZS-CoT	70.3	65.5	69.8	36.0	38.5	21.5	60.4
Sphinx-CCoT	74.2 (+2.6)	68.3 (+2.4)	71.0 (+1.0)	36.5 (+7.5)	36.3 (+7.3)	22.5 (+6.2)	61.9 (+11.9)
GPT4V	69.1	75.5	88.2	60.3	45.3	33.5	64.8
GPT4V-ZS-CoT	72.5	74.8	88.8	63.3	52.5	41.0	65.5
GPT4V-CCoT	74.0 (+4.9)	76.3 (+0.8)	91.2 (+2.0)	64.0 (+3.7)	54.5 (+9.2)	43.3 (+9.8)	67.8 (+3.0)

Table 1. **Main results table on SeedBench, MMBench, Winoground, and WHOOPS! Benchmarks.** Abbreviations: SEEDBench-Image [SEED-I]; Winoground Text Score: Wino-Text, Image Score: Wino-Image, Group Score: Wino-Group. Unlike our zero-shot approach, models with † are supervised and finetuned on annotated scene graphs. For more results, please refer to Section A.2 in Supp.

4.1. Implementation Details

We implemented CCoT using PyTorch [56]. In order to obtain pre-trained models which we evaluated our method, we used each model's respective official implementation. While the compute and memory requirements differ between models, our prompting method needs only the infrastructure necessary for running inference on these models. Refer to Supplementary in Section B for more information.

4.2. Datasets

The goal of our work is to demonstrate that our method improves LMMs' compositional visual understanding, while also enhancing a broad range of vision-and-language tasks. In what follows next, we describe our evaluation datasets.

VL Compositional Benchmarks. To evaluate the compositional visual understanding of our method, we consider the Winoground [68] and WHOOPS! [12] benchmarks: **(1) Winoground** is a hand-curated dataset designed to test VL models' compositional visual understanding. Each sample contains two images and a corresponding pair of image captions. Both captions are syntactically very similar but contain one key difference in the form of a semantic swapping of objects, relations, or both. On the same dataset, Winoground performance is evaluated on three metrics: (i) a text score, where the correct caption must be identified given one image; (ii) an image score, where the correct image must be identified given one caption; (iii) a group score, where the two pairs must be matched correctly. **(2) WHOOPS!** similarly tests compositionality using images that violate typical visual commonsense. There are a broader variety of tasks, in particular: (i) Explanation Generation, (ii) Image Captioning, (iii) Cross-Modal Matching, and (iv) Compositional VQA. We evaluate our method on the Compositional VQA split of the dataset.

Multimodal Reasoning Benchmarks. Recently, there has been an introduction of several new benchmarks that are specifically designed to evaluate the multimodal reasoning abilities of LMMs. In our work, we focus on SEED-Bench [39], MMBench [47], and LLaVA-Bench In-the-Wild [45]. Both SEEDBench and MMBench include different splits that test general visual perception and visual reasoning. For instance, SEEDBench contains perception tasks that evaluate an LMM's Instance Identification and Instance Attribute understanding capabilities while also containing

Model	IC	SU	IId	IA	IL	SR	VR	TU	IIn	Overall
MMCoT†	22.1	29.5	30.2	32.8	33.6	30.3	34.1	45.9	34.0	34.4
LLaVA-1.5-13B-DDCoT	47.3	63.0	59.8	64.1	44.6	41.4	67.1	57.7	51.6	58.0
LLaVA-1.5-13B-VidIL	62.3	74.9	72.5	69.9	62.5	53.9	78.0	49.4	71.1	68.9
LLaVA-1.5-13B-CCoT	59.3	76	74.4	71.8	64.3	54.5	79.2	58.8	74.2	69.7

Table 2. **Comparison to Multimodal CoT Methods.** TBD Instances Counting [IC], Scene Understanding [SU], Instance Identity [IId], Instance Attributes [IA], Instance Location[IL], Spatial Relation [SR], Visual Reasoning [VR], Text Understanding [TU], Instance Interaction[IIn]. Note that †indicates that MMCoT is a finetuning method that was pretrained on ScienceQA.

more higher-order reasoning splits like Scene Understanding and Instance Interaction. MMBench has similar splits. We exclude video, thus evaluating our method on the image splits of SEEDBench and the entirety of MMBench. To evaluate a different type of multimodal reasoning, we further evaluate our method on LLaVA-Bench In-the-Wild, which tests the LMMs' ability to give detailed long-form answers to visual questions.

4.3. Models

In our work, we apply our CCoT approach to four popular LMMs described as follows.

LLaVA-1.5. The LLaVA [46] architecture distinguishes itself as a powerful state-of-the-art (SOTA) LMM method. Featuring a simple linear projection that maps CLIP visual features of the input image into a shared embedding space with the LLM language tokens, LLaVA instruction tunes on a dataset of images–LLaVA-Instruct-158k—paired with conversational, detailed description, and complex reasoning response types for better visual alignment than simply image-text pairs. In our work, we evaluate LLaVA-1.5 [45], a newer version of LLaVA with improved baselines. Model improvements over the original architecture include: (1) replacing the linear projection with an MLP and (2) pretraining on more diverse datasets.

InstructBLIP. While InstructBLIP also uses a frozen visual encoder and LLM, it calculates visual features via a Q-former transformer as in BLIP-2 [41] model that outputs learnable visual tokens. The difference, in this case, is that InstructBLIP's Q-former also attends over the task prompt, making the visual features *instruction-aware*. This, in addition to a broader set of visual instruction tuning datasets that includes the LLaVA-Instruct-158k affords the method high performance on benchmarks like SEEDBench [39].

SPHINX. Sphinx [44] distinguishes itself from other LMMs in two key ways: Sphinx (1) unfreezes its LLM weights during instruction finetuning and (2) has a broader area of multimodal question-answering tasks including "region-level understanding, caption grounding, document layout detection, and human pose estimation" [44].

GPT-4V. Unlike the other three models, GPT-4V's architecture and pretraining details are not made public. However, using the SOTA GPT-4 as the LLM backbone will be essential in evaluating how our method works on an LMM with superior language reasoning skills.

4.4. Baselines

In our experiments, we compare our CCoT prompting methodology to two other prompting baselines as shown in Table 1. First, to evaluate the added benefit of our method to pretrained LMMs, our first baseline is to apply the model to the benchmark without any prompt engineering. Second, we consider a baseline of a *language* zero-shot (ZS) CoT prompting method [36] to determine the benefit of CCoT compared to a SOTA CoT prompting method. The method works in a two-step fashion. (i) Given the input question and text, the reasoning trigger "Let's think step-by-step." is appended to the end of the prompt, coming subsequently after the question. This generates language reasoning for an answer to the question. (ii) Because the answer is implicit in this outputted reasoning, the second step involves passing the image, question, output reasoning from step 1, and an answer extraction phrase to return a response in the desired format. We find that compared to the answer extraction phrase suggested in the original paper, the one suggested by LLaVA [45] yields higher accuracy on most benchmarks and so proceed with this slight change compared to the original implementation of ZS-CoT. We also compare our work to the recent SOTA multimodal CoT prompting methods MMCoT [88], DDCoT [91], and VidIL [74] on the SEEDBench-Image dataset as shown in Table 2

4.5. Results

Results are shown in Table 1. An advantage of our method is that it can be applied across a variety of different pretraining methods and visual architectures. We demonstrate that applying CCoT outperforms the base models across several benchmarks, highlighting the effectiveness of our approach. In Figure 3, we show concrete examples where our method improves upon baselines as well as cases where it still fails. For more results, refer to Section A.2 in Supplementary.

Compositional visual understanding. For all four LMMs tested, we find substantial increases utilizing CCoT compared to baselines when evaluated on Winoground and

Model	SU	IId	IA	IL	SR	VR	IIn	W. Avg.
LLaVA-1.5-13B-CCoT	76.0	74.4	71.8	64.3	54.5	79.2	74.2	72.1
LLaVA-1.5-13B	74.9	71.3	68.9	63.5	51.5	77.0	73.2	69.9
w/ Object Locations	75.4	72.7	69.4	63.6	54.5	78.9	73.2	70.5
w/out JSON Format	74.8	73.1	70.7	63.0	52.0	78.6	73.2	68.1
LLaVA-1.5-13B-Caption-CoT	75.7	73.1	69.1	63.1	55.3	78.6	73.7	70.7
LLaVA-1.5-7B	50.6	42.2	43.0	38.1	33.8	58.0	50.5	66.3
LLaVA-1.5-7B-CCoT	68.7	57.9	63.7	47.9	42.8	67.1	66.0	66.1
128 Token Length	76.2	73.4	71.4	63.7	55.4	80.1	75.3	71.9
512 Token Length	75.5	73.6	71.6	63.2	54.8	79.15	74.2	71.6
1024 Token Length	75.9	73.5	71.7	63.2	54.0	79.5	76.3	71.5

Table 3. **Ablations on SEEDBench-Image.** This table describes key split-level ablation results of our method on all image splits of SEED-Bench [39]: Instances Counting [IC], Scene Understanding [SU],Instance Identity [IIn], Instance Attributes [IA], Instance Location[IL], Spatial Relation [SR], Visual Reasoning [VR], Text Understanding [TU], Instance Interaction[IIn]. W. Avg. denotes the weighted average.

WHOOPS! In fact, without any instruction tuning, GPT-4V-CCoT achieves a significant improvement over the previous Winoground SOTA—SGVL, which has been finetuned on ground truth SG annotations [28]. Interestingly, ZS-CoT method actually *degrades* performances across several splits of the compositional benchmarks. This may be due to the lack of consideration for visual information in the prompt as it was designed for LLMs. Thus, these results demonstrate the effectiveness of CCoT for improving compositional visual reasoning of LMM without the need for finetuning or ground-truth annotated SG data.

Multimodal Benchamarks. We also see that CCoT improves over the baselines on SEEDBench image splits, MMBench, and LLaVA-Bench In-the-Wild. Even with many LMMs having a variety of different LLM backbones and pretraining methods, the difference between consecutive state-of-the-art models on SEEDBench is usually 1% or less. All of CCoT's improvements are 1% or better. Therefore, these results are a robust indication that our method is advantageous for elevating both LMMs' compositional visual understanding and their general multimodal reasoning. Once again, ZS-CoT prompting is actually often detrimental to LMMs on many splits of these benchmarks.

4.6. Ablations

We perform a comprehensive ablation study on SEED-Bench with our LLaVA-1.5-CCoT model (see Table 3). We note that we did not report the Instance Counting and Text Understanding (OCR) splits as they do not constitute visual reasoning. For more ablations, refer to Section A.1 in Supp.

Requiring bounding boxes. In our qualitative exploration of generated SGs, we found that some SGs included bounding-box coordinates for objects. Thus, we experimented with a prompt that instructed the LMM to include bounding-box coordinates (shown in the table as"w\ Object Locations") for all objects in the generated SG. We find a 1.6% decrease in weighted average accuracy on SEEDBench-Image suggesting that requiring exact object locations is not beneficial to multimodal reasoning tasks.

JSON structure enhances SG utilization. While SGs are structured visual representations, they may come in many different textual formats. As such, we ablate the JSON format requirement (refer to as *w/out JSON Format*) of our SG generation prompt to evaluate whether enforcing a specific SG format affects the LMMs usage of the content. Our results indicate that enforcing a common, systematic format like JSON is indeed beneficial (-2.0% without JSON) to the LMMs ability to most effectively utilize the SG.

Replacing SGs with captions. SGs are a *highly-structured* representation of visual information that distinguishes them from simply natural language descriptions of images. Therefore, we ablate the importance of SG structure by generating captions instead of SGs (refer to as *LLaVA-1.5-Caption-CoT*). We find in Table 3 that generating captions with the *same informational context as our SG method*, but, degrades performance (-1.4% compared to ours), suggesting the importance of SG structure to multimodal tasks.

LMM size. We also evaluate the impact of the LMM size. We find that LLaVA-1.5-7B-CCoT shows no noticeable difference (+.1 %) in accuracy compared to LLaVA-1.5-7B. The more substantial gains of LLaVA-1.5-13B-CCoT and GPT-4-CCoT indicate that our method is most effective for larger model sizes. This facet is crucial as our zero-shot method becomes a comparatively less compute-expensive process than the finetuning for these larger LMMs.

Effect of SG size. We consider how the size of the SG affects the generated response, by comparing the accuracy when using SGs of different token lengths. Concretely,

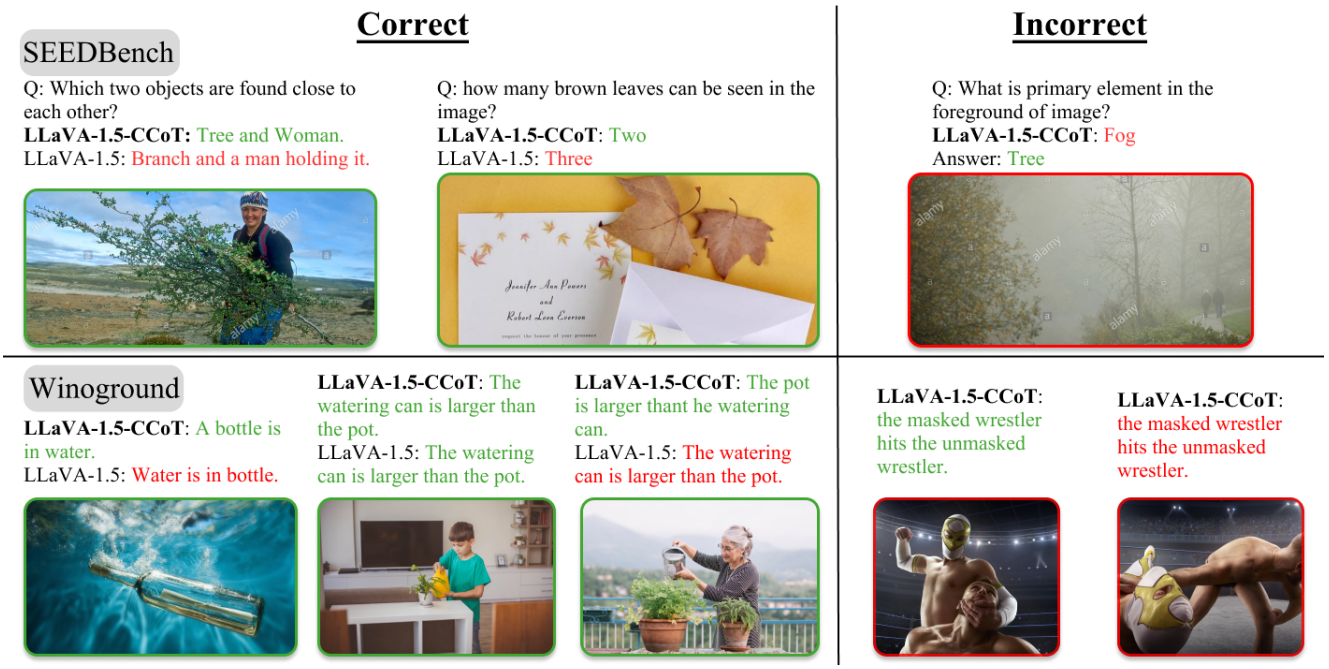

Figure 3. **Example Outputs.** Above we show examples of our method on both SEEDBench and Winoground. On the left we show successes of CCoT while the right shows failure cases. For more qualitative visualizations, please refer to Section C in Supplementary.

we evaluate when using SGs of length 1024 (-.6%), 512 (-0.5%), and 128 (-0.3%) tokens. The results demonstrate that the optimal SG size is 256 tokens. This demonstrates the effectiveness of textual SGs in encapsulating useful information in a small sequence length while also providing credence to the idea that a minimum amount of information is necessary to properly respond to the question.

5. Conclusion

Our research has demonstrated the significant potential of the CCoT approach in extracting compositional information from an LMM. This extracted knowledge leads to enhanced compositional visual and multimodal reasoning of LMMs downstream without the need for fine-tuning or reliance on ground-truth annotated SG data. Our method stands out by generating SGs in a zero-shot manner, effectively addressing the issue of annotated SG availability. Using the generated SG in a CoT reasoning prompt also addresses catastrophic forgetting by not fine-tuning. The substantial improvements observed on compositional visual reasoning benchmarks like Winoground and WHOOPS!, along with the general multimodal benchmarks SEEDBench, MM-Bench, and LLaVA-Bench In-the-Wild underscore the effectiveness of our approach across a diverse set of tasks. This is further corroborated by our ablations, which reveal the importance of using structured SGs over captions, leveraging the JSON format, and utilizing optimal SG length to enhance the LMMs' visual compositional and multimodal reasoning. These results collectively highlight the value of our method in broadening the capabilities of LMMs in compositional and multimodal reasoning tasks.

6. Limitations

In this work, we present a zero-shot Chain-of-Thought prompting method that utilizes scene-graph representations for multimodal and compositional visual reasoning tasks. We demonstrate improved performance on several different models and benchmarks. Nevertheless, our work has a central limitation. While extending context length is an active field of research, our method is limited by the current context lengths of the LLMs being used by the LMMs. Additionally, scene graphs are not particularly useful representations when performing multimodal tasks that emphasize language over visual reasoning, such as document understanding. Finally, we do not anticipate negative impacts of this work, but, as with any machine learning method, we recommend exercising caution.

Acknowledgements.

We would like to thank Suzie Petryk, Alon Mendelson, Sanjay Subramanian, Rudy Corona, and Leonid Karlinsky for helpful feedback and discussions. This project was supported in part by DoD, including PTG and/or LwLL programs, as well as BAIR's industrial alliance programs.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 2

[2] Amit Alfassy, Assaf Arbelle, Oshri Halimi, Sivan Harary, Roei Herzig, Eli Schwartz, Rameswar Panda, Michele Dolfi, Christoph Auer, Peter W. J. Staar, Kate Saenko, Rogerio Feris, and Leonid Karlinsky. FETA: Towards specializing foundational models for expert task applications. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. *ArXiv*, abs/1511.02799, 2015. 2

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2015. 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 3

[6] Elad Ben Avraham, Roei Herzig, Karttikeya Mangalam, Amir Bar, Anna Rohrbach, Leonid Karlinsky, Trevor Darrell, and Amir Globerson. Bringing image scene structure to video via frame-clip consistency of object tokens. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 3

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023. 2

[8] Amir Bar, Roei Herzig, Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor Darrell, and A. Globerson. Compositional video synthesis with action graphs. In *ICML*, 2021. 3

[9] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *ECCV*, pages 105–121, 2018. 3

[10] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 3

[11] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *ArXiv*, abs/2308.09687, 2023. 2

[12] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *ArXiv*, abs/2303.07274, 2023. 5, 4

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022. 2

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2, 4

[17] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. 2022. 2

[18] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *2023 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 2657–2668, 2022. 1, 3

[19] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. 2

[20] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022. 2

[21] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qianmengke Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *ArXiv*, abs/2305.04790, 2023. 2

[22] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962, 2022. 2

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. 2

[24] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 3

[25] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-temporal action graph networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[26] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. Learning canonical representations for scene graph to image generation. In *European Conference on Computer Vision*, 2020. 3

[27] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[28] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision \& language models using scene graphs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 1, 2, 3, 7

[29] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019. 2, 3, 1

[30] Achiya Jerbi, Roei Herzig, Jonathan Berant, Gal Chechik, and Amir Globerson. Learning object detection from captions via textual scene attributes. *ArXiv*, abs/2009.14558, 2020. 3

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[32] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 3

[33] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3008–3017, 2017. 2

[34] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. 2

[36] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. 2, 6

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2, 3, 1

[38] Bin Lei, Pei-Hung Lin, Chunhua Liao, and Caiwen Ding. Boosting logical reasoning in large language models through a new framework: The graph of thought. *ArXiv*, abs/2308.08614, 2023. 2

[39] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023. 1, 5, 6, 7, 2

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 6

[42] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 3

[43] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 3

[44] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Jiao Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ArXiv*, abs/2311.07575, 2023. 4, 6

[45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 4, 5, 6

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 6

[47] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. 1, 5, 3

[48] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513, 2022. 1

[49] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *ArXiv*, abs/2304.09842, 2023. 2

[50] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, H. Fu, Qinghua Hu, and Bing Wu. Fairness-guided few-shot prompting for large language models. *ArXiv*, abs/2303.13217, 2023. 2

[51] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *ArXiv*, abs/2212.07796, 2022. 1, 3

[52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. 1, 2, 3

[53] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[54] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv*, abs/2202.12837, 2022. 2

[55] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1, 2, 4

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[57] Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. *ArXiv*, abs/2307.16789, 2023. 2

[58] Moshiko Raboh, Roei Herzig, Gal Chechik, Jonathan Berant, and Amir Globerson. Differentiable scene graphs. In *WACV*, 2020. 3

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[60] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. 2

[61] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23:289 – 301, 2022. 2

[62] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761, 2023. 2

[63] Yongliang Shen, Kaitao Song, Xu Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580, 2023. 2

[64] Sanjay Subramanian, Medhini G. Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. Modular visual question answering via code generation. *ArXiv*, abs/2306.05392, 2023. 2

[65] D'idac Sur'is, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *ArXiv*, abs/2303.08128, 2023. 2

[66] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. pages 5099–5110. Association for Computational Linguistics, 2019. 3

[67] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. Ul2: Unifying language learning paradigms. In *International Conference on Learning Representations*, 2022. 2

[68] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 5238–5248, 2022. 3, 5, 4

[69] Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Ö. Arik, and Tomas Pfister. Better zero-shot reasoning with self-adaptive prompting. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 2

[70] Lei Wang, Yilang Hu, Jiabang He, Xingdong Xu, Ning Liu, Hui juan Liu, and Hengtao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *ArXiv*, abs/2305.03453, 2023. 3

[71] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 2

[72] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 3

[73] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022. 2

[74] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. Language models with image descriptors are strong few-shot video-language learners. *ArXiv*, abs/2205.10747, 2022. 2, 6

[75] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021. 2

[76] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 2, 1

[77] Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv*, abs/2303.04671, 2023. 2

[78] Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts. *ArXiv*, abs/2305.14688, 2023. 2

[79] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *CVPR*, pages 3097–3106, 2017. 2, 3

[80] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, 2022. 2

[81] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601, 2023. 2

[82] Yao Yao, Z. Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *ArXiv*, abs/2305.16582, 2023. 2

[83] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 2, 3

[84] Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257, 2023. 2

[85] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216, 2021. 3

[86] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 3

[87] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alexander J. Smola. Automatic chain of thought prompting in large language models. *ArXiv*, abs/2210.03493, 2022. 2

[88] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander J. Smola. Multimodal chain-of-thought reasoning in language models. *ArXiv*, abs/2302.00923, 2023. 3, 6

[89] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *ArXiv*, abs/2307.04087, 2023. 2

[90] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 3, 1

[91] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *ArXiv*, abs/2310.16436, 2023. 2, 6, 1

[92] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2