Compositional Chain-of-Thought Prompting for Large Multimodal Models

Supplementary Material

Here we provide additional information about our experimental results, qualitative examples, implementation details, and datasets. Specifically, Section A provides more experiment results, Section B provides additional implementation details, and Section C provides qualitative visualizations to illustrate our approach.

A. Additional Experiment Results

We begin by presenting several additional ablations (Section A.1) that further demonstrate the benefits of our CCoT approach. We also present additional results (Section A.2) on MMBench Perception Splits.

A.1. Additional Ablations

In what follows, we provide additional ablations that further illustrate the benefits of CCoT. For all ablations, we compare the ablated experiment with the corresponding best-performing CCoT results on the SEEDBench-Image dataset.

Random SG Regularization. To assess whether only the structure of SGs can be valuable for reasoning, we ablate the specific details of the scene graph. Concretely, we pass a random SG not related to the input question to the LMM and prompt the model to only use it as a framework for reasoning, essentially regularizing the response via the random SG. We find a 5.6% degradation in performance from **LLaVA-1.5-CCoT** as shown in Table 4. The ablation demonstrates that providing the structure of an SG without accurate content leads to an accuracy decrease compared to LLaVA-1.5-CCoT (-5.6%). The capability to generate correct answers despite suboptimal reasoning steps is supported by recent literature—e.g. CoT [76] and DDCoT [91]. This demonstrates that our method does not require ground truth SGs but also generates SGs accurately enough to make LMMs substantially more effective on compositional & multimodal tasks.

SG Knowledge Distillation. Here, we ask whether the compositional knowledge extracted in high-quality scene graphs can be transferred to other models. In particular, we use **LLaVA-1.5-13B-CCoT** scene graphs when prompting InstructBLIP on SEEDBench-Image splits. The entire process for running **InstructBLIP-13B-CCoT** remains the same with just the scene graph being replaced by the one generated by **LLaVA-1.5-13B-CCoT**. We find that this actually leads to a slight degradation in performance of **InstructBLIP-13B-CCoT**, but still better than InstructBLIP-ZS-CoT.

Removing Image. Finally, we ablate the effect of removing the image during the second response generation step. The first step of our method is left intact, using the image and question as context to yield a relevant scene graph. However, in the second step, we generate a response with the image tokens masked. This experiment evaluates the gap between the SG and the visual information offered by the image. We find a significant decrease in performance of 26.1%, indicating that it is a combination of the image and scene graph that affords our method its improved performance over baselines.

Impact of individual compositional characteristics. We individually ablate the objects, attributes, and relationships of the generated CCoT SGs. We find improvements in accuracy over LLaVA-1.5 when we remove the objects/attributes/relationships from LLaVA-1.5-13B-CCoT by 1.5/0.8/1.2 % on SEEDBench and 3.5/3.5/2.25 % on Winoground Text. These results indicate that each of the three elements of CCoT contributes to its effectiveness, but combined, they achieve the best result.

Removing the term ”Scene Graph” from prompt. It is possible that the term ”scene graph” encapsulates some latent understanding of visual compositionality that the LMM has already. To evaluate this, we ablate the term ”scene graph” from our prompt and simply replace it with the word ”description”. Indeed this shows a -2.1% decrease in accuracy on the evaluated SEEDBench-Image splits. This suggests that LMMs have some knowledge of SGs from pretraining that is helpful when generating a structured compositional representation to aid in multimodal reasoning.

COCO and Visual Genome Data Overfitting. Most LMMs have been extensively instruction-tuned on images sourced from the COCO [43] and Visual Genome [37] datasets. In fact, LLaVA-Instruct-158k, the main dataset used to finetune LLaVA and many other LMMs consists solely of COCO images in order to make use of the bounding box and detailed description information provided by the dataset. Besides this, there are many instances of other datasets that reuse COCO and Visual Genome images. As such, our method is effective in helping these LMMs generalize to datasets like SEEDBench, MMBench, and Winground, while showing no substantial difference in performance compared to the zero-shot case on benchmarks like GQA [29] and VL Checklist [90] that heavily make use of the overfit COCO or Visual Genome images (CCoT is -.8 % and -.6 % compared to zero-shot on GQA and VL Checklist respectively).

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/ Random Scene Graphs	73.4	71.3	67.2	62.2	50.2	77.3	75.3	66.5
w/out Objects	76.0	73.8	71.6	63.4	52.3	79.5	76.3	71.4
w/out Attributes	75.7	73.5	71.2	63.9	52.5	79.2	72.2	70.7
w/out Relationships	75.4	73.1	71.8	64.3	52.8	79.5	74.2	71.1
w/out Image	49.2	46.6	47.1	43.2	38.5	54.7	50.5	46.0
w/out "Scene Graph"	74.7	72.3	72.5	60.4	53.0	77.0	72.2	69.8
InstructBLIP-13B-CCoT	68.7	57.9	63.7	47.9	42.8	67.1	66.0	60.1
w/ LLaVA-1.5-13B-CCoT SGs	50.6	42.2	43.0	38.1	33.8	58.0	50.5	44.8

Table 4. **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.

Model	IC	SU	IId	IA	IL	SR	VR	TU	IIn
InstructBLIP-13B	29.7	60.3	55.4	51.0	41.8	32.4	46.8	31.8	47.42
InstructBLIP-13B-CCoT	34.2	68.7	57.9	63.7	47.9	42.8	67.1	40.0	66.0
LLaVA-1.5-13B	61.3	74.9	71.3	68.9	63.5	51.5	77.04	60	73.2
LLaVA-1.5-13B-CCoT	59.3	76	74.4	71.8	64.3	54.5	79.2	58.8	74.2

Table 5. **Detailed Results Table SEEDBench.** This table describes the split-level results of our method on all image splits of SEED-Bench [39]: 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]].

A.2. Additional Results

Detailed Split Results. We present detailed results of our method on the individual SEEDBench-Image splits as well as all of MMBench's splits which are separated into their Perception and Reasoning categories. These results shown in Tables 5, 6, and 7.

LLaVA-Bench Results. LLaVA-Bench is a challenging hand-designed dataset on a small number of images and questions for evaluating the effectiveness of LMMs as multimodal chatbots. The questions are designed in a more open-ended conversational manner, differing greatly from the other benchmarks presented in the main paper. Instead of evaluating a simple response to a question, the benchmark tests the entire long-form text conversions, meaning there is a greater burden to account for the language response. Examples of LLaVA-Bench are shown in Figure 3. Even so, we show slight improvements over the baselines– +0.4% for InstructBLIP and +0.3% for LLaVA-1.5–when ZS-CoT shows significant degradation (roughly 10% decrease). This suggests that our method is also potentially beneficial for more open-ended visual chat applications, which is different than multimodal reasoning and VL compositional benchmarks.

B. Additional Implementation Details

To run our models on larger benchmarks, we use 8 NVIDIA RTX 6000 GPUs to split the datasets across multiple GPUs. Smaller datasets like Winoground and MM-Vet are able to run individual experiments on a single GPU. Besides the output token generation length, we use the default generation parameters (e.g. temperature and no. of beams in beam search) for each model. For any baseline performance already reported by the official codebase of the model (e.g. LLaVA-1.5 on SEEDBench or MMBench), we use the value presented in that model's corresponding paper. Please refer to the respective model's paper for their specific implementation details of the architecture. In the following sections, we describe some nuances of our method applied on different datasets.

B.1. SEEDBench

Dataset. SEED-Bench [39]is a large-scale benchmark designed to provide a comprehensive and objective evaluation of LMMs, particularly focusing on generative comprehension. This benchmark comprises 19k multiple-choice questions, all of which have been annotated by humans. These questions are structured to cover 12 evaluation dimensions,

Model	LR	AR	RR	FP-S	FP-C	CP
InstructBLIP-13B	11.5	43.6	35.5	36.6	22.3	51.7
InstructBLIP-13B-CCoT	12.5	45.8	40.9	40.7	22.1	56.0
LLaVA-1.5-13B	39.9	74.7	61.6	70.9	59.9	75.4
LLaVA-1.5-13B-CCoT	44.2	72.1	75.3	73.7	59.3	81.2

Table 6. **Detailed Results Table MMBench Reasoning.** This table describes the split-level results of our method on splits classified as Reasoning by MMBench [47]: Logic Reasoning [LR], Attribute Reasoning [AR], Relation Reasoning [RR], Fine-Grained(Single) [FG-S], Fine-Grained (Cross) [FG-C], Coarse Perception [CP].

	Coarse Perception					FGSI				FGCI	
Model	IT	IQ	IE	IS	IS	OCR	CR	OL	ARS	AC	SR
InstructBLIP-13B	16.7	14.8	50.0	37.1	22.6	35.0	53.5	4.9	7.1	2.1	1.0
InstructBLIP-13B-ZS-CoT	16.7	3.0	36.0	36.1	20.8	37.5	39.3	6.17	40.4	2.1	2.2
InstructBLIP-13B-CCoT	61.1	9.3	54.0	82.9	49.1	45.0	60.0	8.64	45.8	11.4	4.4
LLaVA-1.5-13B	83.3	50.0	86.0	95.2	73.6	57.5	81.8	45.7	87.0	61.4	93.0
LLaVA-1.5-13B-ZS-CoT	80.5	55.6	82.0	95.2	81.1	57.5	78.8	40.7	92.2	59.1	26.7
LLaVA-1.5-13B-CCoT	81.5	44.4	86.0	97.1	83.0	62.5	84.8	53.1	87.0	83.9	31.1

Table 7. **Detailed Results Table MMBench Perception.** This table describes the split-level results of our method on splits classified as Reasoning by MMBench[]. Category Abbreviations:Fine-Grained Perception (Single-Instance) [FGSI], Fine-Grained Perception (Cross-Instance) [FGCI]; Split Abbreviations: Image Topic [IT], Image Quality [IQ], Image Emotion [IE], Image Scene [IS], Image Style [IS], OCR [OCR], Celebrity Recognition [CR], Object Localization [OL], Attribute Recognition (Single-Instance) [ARS], Attribute Recognition (Cross-Instance) [ARC] Attribute Comparison [AC], Spatial Relationship [SR].

catering to both spatial and temporal understanding.

The development of SEED-Bench involved designing an advanced pipeline specifically for creating multiple-choice questions. This pipeline is tailored to target specific evaluation dimensions, thereby enabling the scalability of evaluation data across various domains. Furthermore, the benchmark incorporates a blend of automatic filtering and manual verification processes to enhance the quality of the generated questions and answers.

For the specific research paper in question, only the image splits of SEED-Bench are utilized for evaluation purposes.

Inference details. We use the official dataset released by the authors which is available at https://github.com/AILab-CVC/SEED-Bench. All models evaluated on SEEDBench use the exact method described in Section 3 of the main paper.

B.2. MMBench

Dataset. MMBench [47] is a novel multimodal benchmark created to address the limitations of existing benchmarks like VQAv2 [5] or COCO Captions[43], which provide quantitative performance measurements but lack fine-grained ability assessment and robust evaluation metrics. Unlike subjective benchmarks such as OwlEval[83], which offer comprehensive evaluations of a model's abilities through human labor but suffer from scalability issues and bias, MMBench offers a more systematic and objective approach.

MMBench consists of two primary components: (i) Curated Dataset. MMBench features a dataset that stands out from existing benchmarks in terms of the number and variety of evaluation questions and abilities (e.g. splits that evaluate emotion or celebrity recognition to test outside knowledge of an LMM). (ii) CircularEval Strategy with ChatGPT Integration. The second key component of MMBench is the CircularEval strategy, which is complemented by the integration of ChatGPT. This approach is aimed at converting free-form predictions into predefined choices, leading to a more robust and reliable evaluation of the LMMs' predictions.

MMBench is thus a comprehensive evaluation pipeline that improves upon existing benchmarks in both scale and depth of assessment.

Inference details. We use the official data and code released by the authors which is available at https://github.com/open-compass/MMBench. All models evaluated on MMBench use the exact method described in Section 3 of the main paper. Final results are obtained by submitting the output predictions to the official MMBench scoring system at https://mmbench.opencompass.org.cn/mmbench-submission

B.3. Winoground

Dataset. Winoground[68] is designed to evaluate the compositional understanding of vision-and-language (VL) models. It challenges these models to correctly pair text and images that share the same underlying compositional structure but differ in the objects involved. Winoground provides a way to assess whether models truly understand the composition of scenes and descriptions, or if they are merely exploiting superficial correlations in the training data.

The Winoground benchmark consists of 400 sets of images and captions. Each set includes two images and two captions, where each image corresponds to one of the captions. The images and captions are carefully designed to be compositionally similar but involve different objects. For example, a set might include an image of a cat chasing a dog with a corresponding caption, and another image of a dog chasing a cat with its caption.

The benchmark evaluates VL models on three scores: (1) Text Score. This score assesses the model's ability to match text captions to the correct images. A high text score indicates that the model effectively understands and applies compositional structures in language. (2) Image Score. This score evaluates how well the model matches images to the corresponding text captions. A high image score suggests a strong understanding of compositional structures in visual data. (3) Group Score. This score is the average of the text and image scores. It provides a holistic measure of the model's overall performance in understanding and applying compositional structures across both visual and textual data.

The Winoground benchmark is significant because it moves beyond traditional benchmarks that often allow models to succeed by leveraging simple heuristics or biases in the data. Instead, Winoground requires models to demonstrate a genuine understanding of the compositional relationships between objects in images and their descriptions, demonstrating our model's value in enhancing compositional visual reasoning in LMMs.

Inference details. We use the official data released by the authors which is available at `https://huggingface.co/datasets/facebook/winoground`. Since we evaluate on LMM methods that were designed for single-image inference, we perform a two-step answer extraction process for the image and group tasks (which have two images). (1) First, instead of asking for an answer-choice as in multiple-choice formatted questions, we ask the LMM to generate reasoning for each image caption pair. (2) Secondly, both text reasoning responses are prompted to GPT-4 to yield the LMM's intended answer.

B.4. WHOOPS!

Dataset. The WHOOPS![12] dataset is a distinctive benchmark developed to evaluate AI models' visual commonsense reasoning, with a particular emphasis on compositional understanding. It consists of 500 synthetic images, each uniquely designed to defy commonsense norms, accompanied by 10874 annotations. These images, crafted using advanced text-to-image models such as Midjourney, DALL-E, and Stable-Diffusion, present scenarios that are logically or physically implausible, thus challenging AI models to go beyond simple object recognition and delve into deeper interpretative reasoning.

The dataset is notable for its diverse array of 'weirdness' categories, encompassing temporal discrepancies, biological rules, cultural knowledge, and more. Each image in WHOOPS! is an invitation for AI models to engage in sophisticated multi-step reasoning, connecting visual cues to knowledge about the world in ways that require a nuanced understanding of everyday experiences, physical and social knowledge, and cultural norms.

WHOOPS! offers four distinct tasks for model evaluation: (i) Explanation Generation. Where models must articulate detailed reasons behind the unusual nature of an image. Image Captioning: Involving the summarization of the images' content. Cross-Modal Matching: Requiring models to differentiate between detailed and underspecified captions. (ii) Visual Question Answering (VQA). This task specifically assesses the models' ability to understand and interpret compositional elements in the images. In the context of the research paper, the Visual Question Answering (VQA) task of the WHOOPS! dataset was chosen for evaluation. This task is designed to test models' compositional understanding and reasoning. It requires models to answer questions that probe their comprehension of the unusual or 'weird' elements within the images, focusing on their ability to integrate visual information with commonsense knowledge. This task is particularly relevant for assessing how well AI models grasp the implausible or unconventional contexts in which objects are depicted, demanding an advanced level of compositional reasoning. By selecting the VQA task from the WHOOPS! dataset, our work aims to critically evaluate and advance the capabilities of LMM models in compositional visual understanding.

Inference details. We use the official data released by the authors which is available at `https://whoops-benchmark.github.io/`. For our evaluation on the VQA split, we use the same answer extraction and evaluation process as the paper [12].

B.5. LLaVA-Bench (In-the-Wild)

Dataset. LLaVA-Bench (In-the-Wild)[45] is a newly developed benchmark that has been used to evaluate the ability of LMMs to provide detailed, yet generalized chat responses to multimodal questions on a variety of images. Given an image, the LMM is prompted with a multimodal task. The LMM's response is compared to GPT-4 generated responses to assess the quality of the response. Although still in development, this small, hand-designed benchmark demonstrates the effectiveness of our method on multimodal chat scenarios.

Inference details. We use the official data released by the authors which is available at https://github.com/haotian-liu/LLaVA/blob/main/docs/LLaVA_Bench.md. For this more open-ended task, we first generate the zero-shot response. Following this, we use our method to generate a scene-graph and then improve the original response. This is to account for the fact that the dataset resembles a long-form conversation.

C. Qualitative Visualizations

We present further qualitative success and failure cases of **LLaVA-1.5-CCoT** in Figure 4.

D. Licenses and Privacy

The license, PII, and consent details of each dataset are in the respective papers. In addition, we wish to emphasize that the datasets we use do not contain any harmful or offensive content, as many other papers in the field also use them. Thus, we do not anticipate a specific negative impact, but, as with any machine learning method, we recommend exercising caution.

Correct	Incorrect

MMBench

Q: What is true about the hurricane?
Choice 1: Can be found only over water.
Choice 2: large spiral-shaped storms.
Choice 3: Can be found over land.
A: Hurricane are large spiral-shaped storms.

Q: Are the two horses in the picture the same size?
Choice 1: Same.
Choice 2: Not the same.
Choice 3: Can't judge.
A: Not the same.

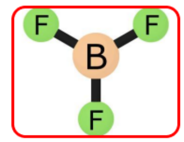
Q: Which property do these two object have in common?
Choice 1: Salty. Choice 2: Yellow.
A: Salty

Q: Which animal's skin is also adapted for survival in cold places?
Choice 1: fantastic leaf-tailed gecko.
Choice 2: Polar bear.
A: Polar bear

Q: Which of is one characteristic of tropical coral reef?
Choice 1: Usually found in deep ocean.
Choice 2: Used by many different organism
Choice 3: Have many large rocks called coral.
A: Used by dufferent organism

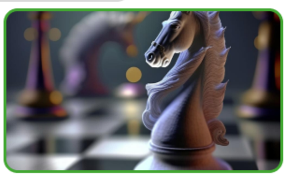
Q: What will happen next?
Choice 1: This person will cry.
Choice 2: This person will mad.
Choice 3: This person will laugh.
Choice 4: Both 1,2,3
A: This person will laugh

Q: Which material is this spatula made of?
Choice 1: Robber. Choice 2: Cotton
A: Cotton



Q: Complete sentence. Boron Trifluorides is
Choice 1: elementary substance
Choice 2: compound
A: elementary substance

WHOOPS!



Q: What is the pictured with a unicorn chess piece?
A: Chess Board

Q: What brand of Polo does Sundar Pichai wear?
A: OpenAI

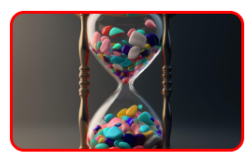
Q: How much chocolate cake is included?
A: One piece

Q: What grows out of an icey pool of water?
A: Cactus

Q: Who wears a hat with google logo?
A: Man

Q: Where are the car competing?
A: Roman Colosseum

Q: A green stop sign stands next to what?
A: Building



Q: What is filled with the colorful chewing gum?
A: Clock

LLaVA-Bench Examples

Q: Describe the photo in detail.
A: The image features a beautiful beach with a large mountain in the background. The mountain is covered in trees, and it appears to be a volcano. The beach is situated at the base of the mountain, providing a picturesque view of the landscape. \n\nIn the distance, there are several houses and buildings scattered around the area, likely part of a city or town. The houses are of various sizes and are spread out across the scene. The combination of the beach, mountain, and city creates a unique and captivating environment.

Q: Discuss how this creative twist on a classic work of art might be interpreted differently by various audiences.
A:The creative twist of painting a dog to resemble the Mona Lisa, a famous painting by Leonardo da Vinci, can be interpreted differently by various audiences. Some people might find it amusing and creative, as it combines two unrelated subjects in a unique and unexpected way. This can be seen as a form of artistic expression and a playful take on a classic work of art.\n\nOn the other hand, some individuals might view this as a disrespectful or inappropriate representation of the Mona Lisa. They might...

Q: What is unusual about this image?
A: The unusual aspect of this image is that a man is ironing clothes while standing on the back of a moving car. This is not a typical scene, as ironing clothes is usually done indoors, in a stationary position, and not on the back of a moving vehicle. The man's actions are not only unconventional but also potentially dangerous, as standing on the back of a car while it is in motion can lead to accidents or injuries.

Q: What are the animals in the painting and what are they doing?
A: The animals in the painting are a bear, a cat, and a rabbit. They are sitting around a small wooden table, eating food and drinking from cups.

Figure 4. **Additional Example Outputs.** Above we show some additional examples of our method on both MMBench, WHOOPS!, and LLaVA-Bench.