Compositional Chain-of-Thought Prompting for Large Multimodal Models

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

关键词： Scene Graph，Chain-of-Thought，MLLM

曹耘宁
2024/10/08
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

研究者

Trevor Darrell

Professor of Computer Science, U.C. Berkeley
Verified email at eecs.berkeley.edu - Homepage

Computer Vision Artificial Intelligence AI Machine Learning Deep Learning

Chang Wen Chen

Chair Professor of Visual Computing, The Hong Kong Polytechnic University
Verified email at polyu.edu.hk - Homepage

multimedia communication multimedia systems image/video processing
multimedia signal processing

Tat-Seng Chua

National University of Singapore
Verified email at comp.nus.edu.sg - Homepage

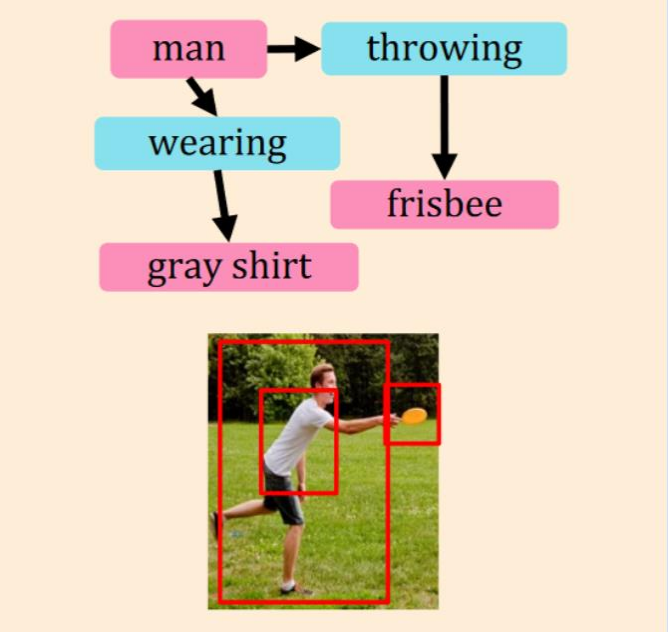
Multimedia Information Retr… Live Social Media Analysis

- 研究背景
- 研究方法
- 实验效果
- 总结

Compositional

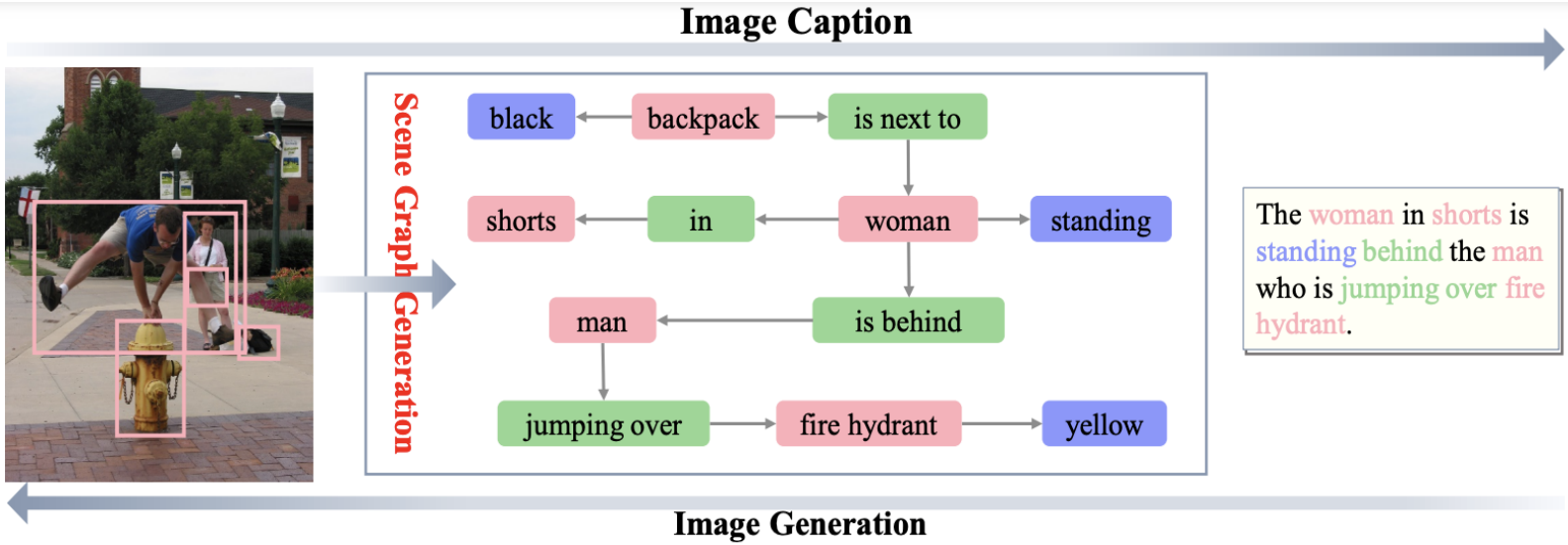
- 什么是组合性？
  - 组合性：总体的含义由它的各个组成部分的含义决定
  - Language：句子是总体，单词是部分
  - Vision：场景是总体，物体、属性及其关系是部分

- 组合理解
  - 物体
  - 物体的属性
  - 物体间的位置关系、动作关系



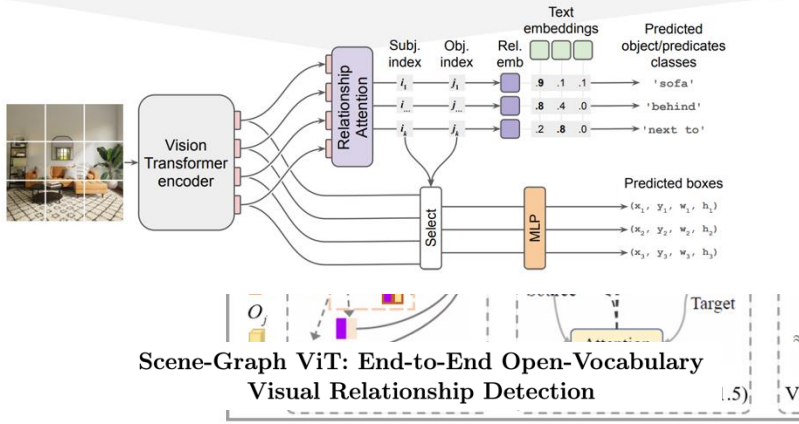
场景图（scene graph）

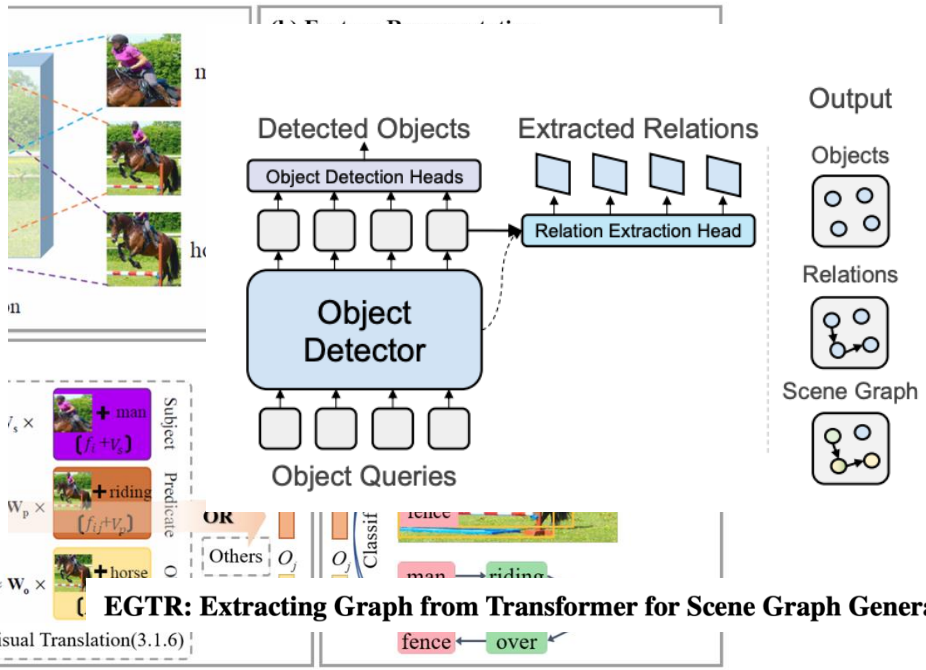
- 什么是场景图?
  - 一种结构化表示（object，attribute，relation）
  - 可作为图-文模态之间的桥梁
- 场景图生成-->用于内容理解



场景图生成（scene graph generation）

- 基于目标检测的方法（image→scene graph）
  - 使用Faster-rcnn，detr框架，额外增加关系预测模块
  - 位置检测精准，但是框架复杂（detector，decoder，relation prediction），发展方向是更简洁的模型设计、端到端框架

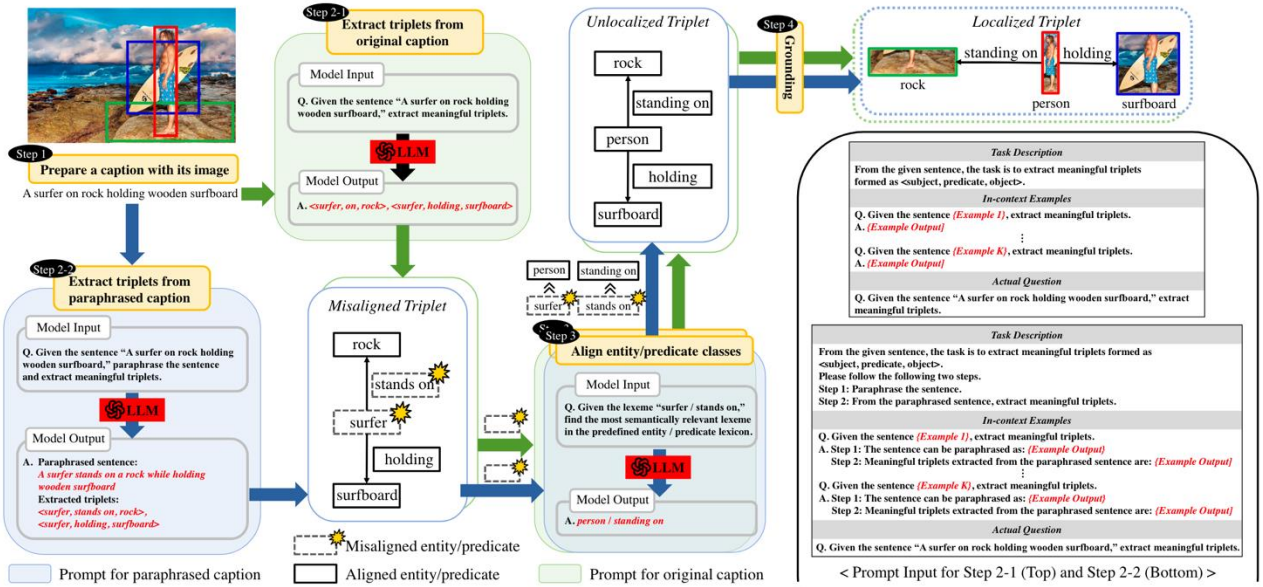




场景图生成（scene graph generation）

- 基于LLM的方法（image+text→scene graph）
  - 引入caption信息预测场景图，发展方向是弱监督、开放词汇

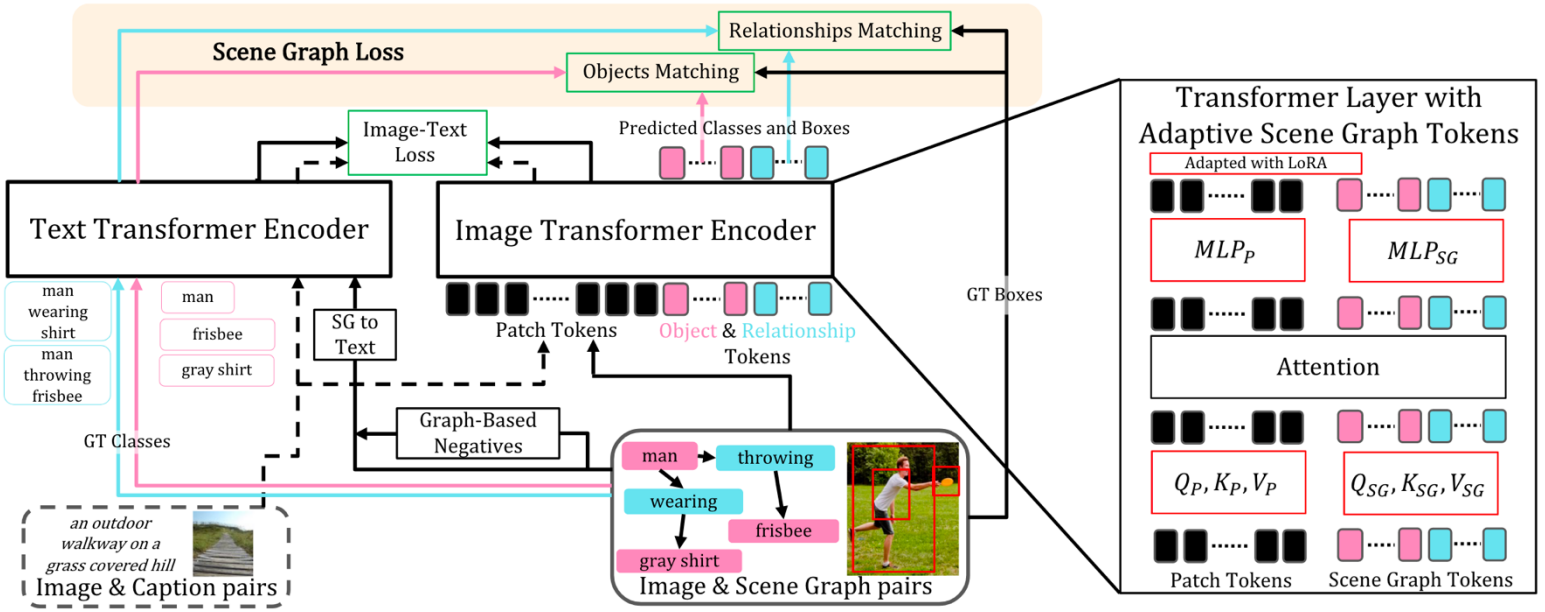
1. LLM从caption、转写的caption中分别提取三元组
2. 三元组对齐
3. grounding



LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation

场景图辅助内容理解

- 在finetune过程中引入场景图信息
  - 标注昂贵、灾难性遗忘
- 使用标注数据



场景图辅助内容理解

- 在finetune过程中引入场景图信息
  - 标注昂贵、灾难性遗忘
- Generation→understanding

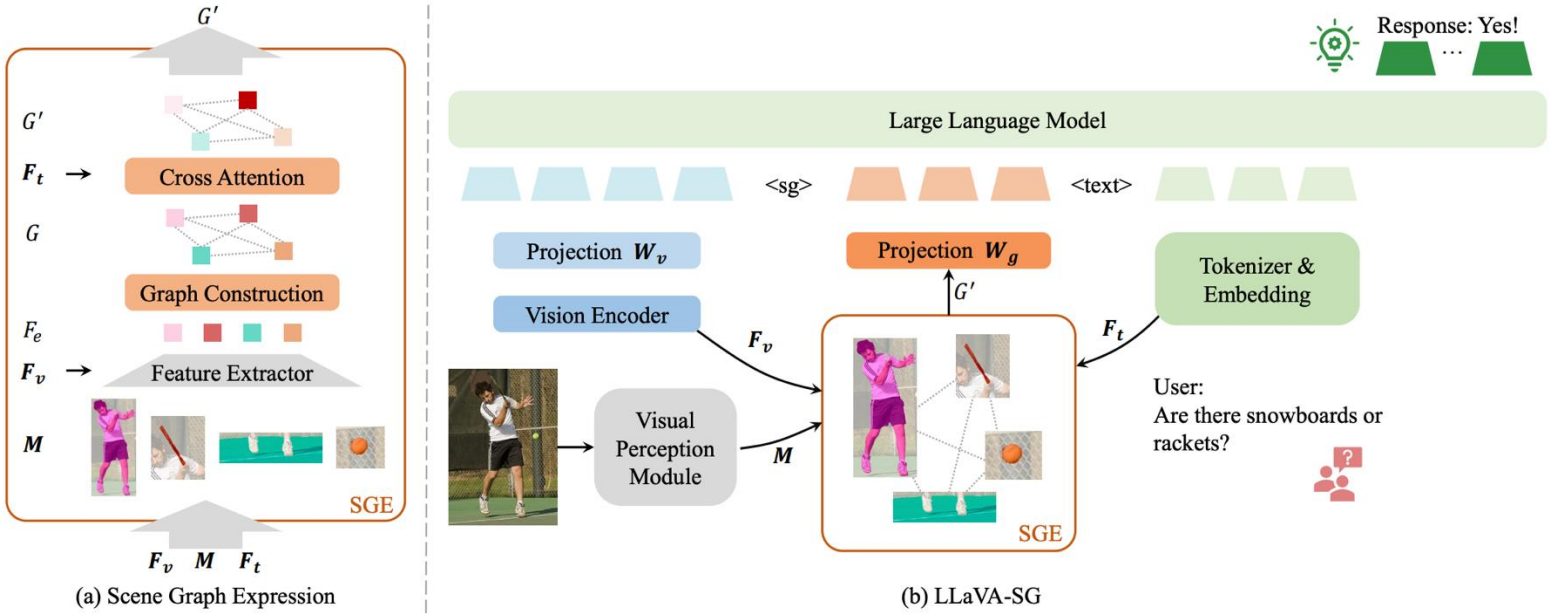


Fig. 2. The structure of the proposed Scene Graph Expression module and the LLaVA-SG framework.

LLaVA-SG: Leveraging Scene Graphs as Visual Semantic Expression in Vision-Language Models

思维链（chain-of-thought）

- 什么是思维链？多步的prompt

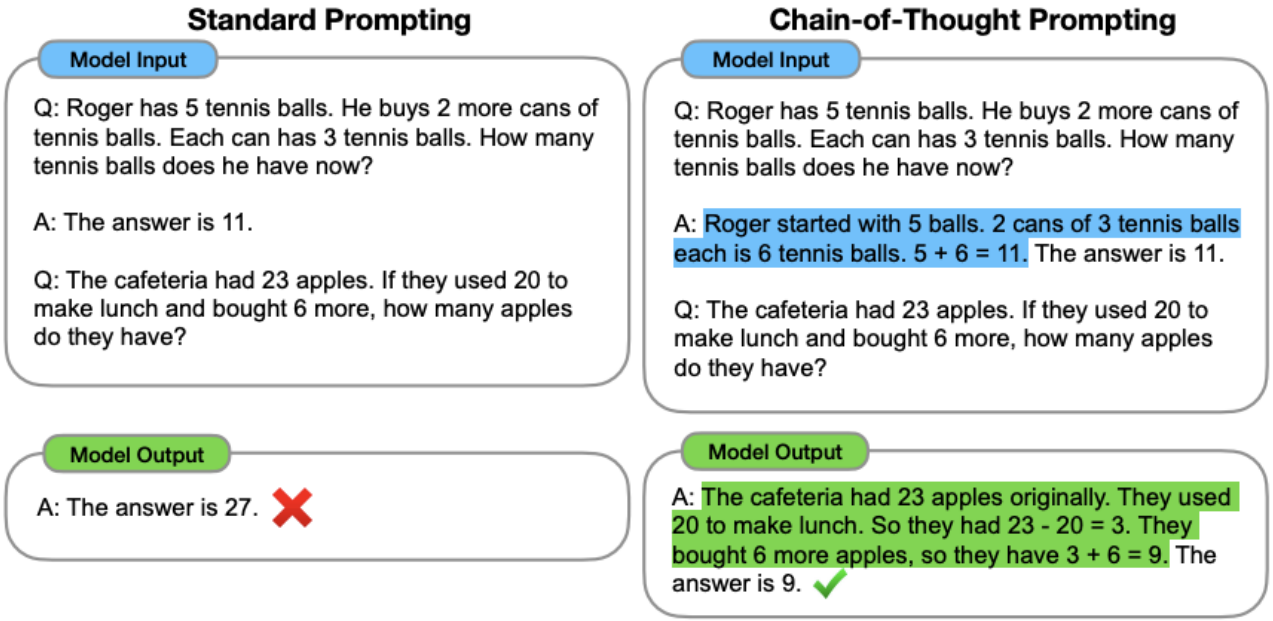
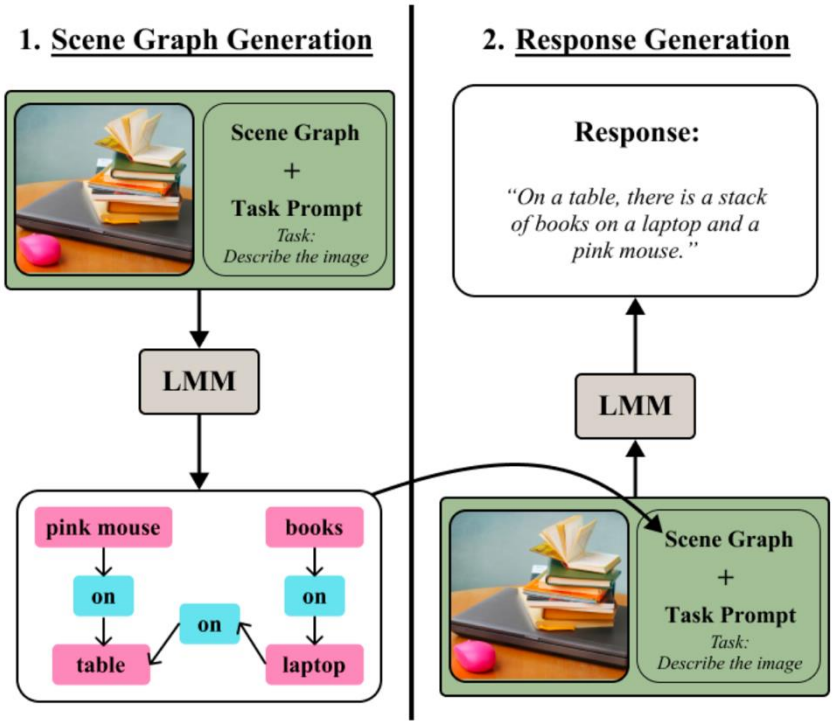


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

研究动机

- Scene graph需要昂贵的标注
- Finetune过程中加入scene graph数据容易导致灾难性遗忘
- 利用思维链提示，引导场景图生成→ 增强下游任务。不需要训练。



- 研究背景
- 研究方法
- 实验效果
- 总结

总体框架

- Chain of thought 框架

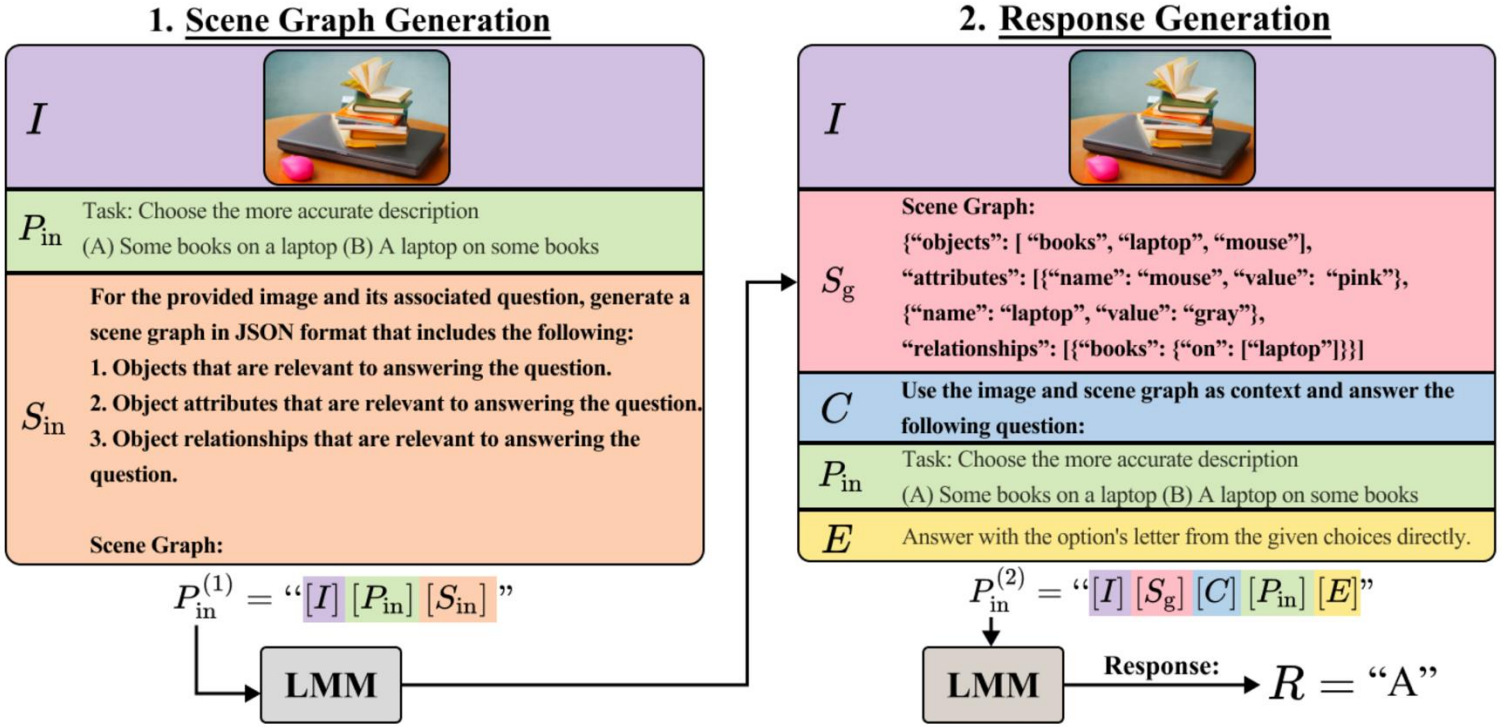
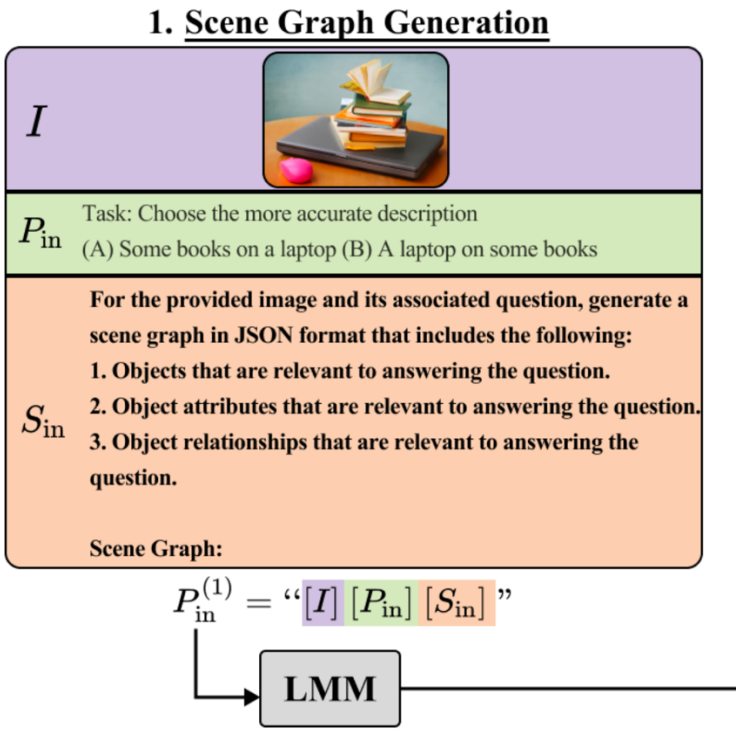


Figure 2. **Full prompt example of CCoT.** The first step in our prompting method is to generate a scene graph given both the image *and* textual task as context. Following this, the answer is extracted by prompting the LMM with the image, scene graph, question, and answer extraction prompt. Prompt sections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

场景图生成

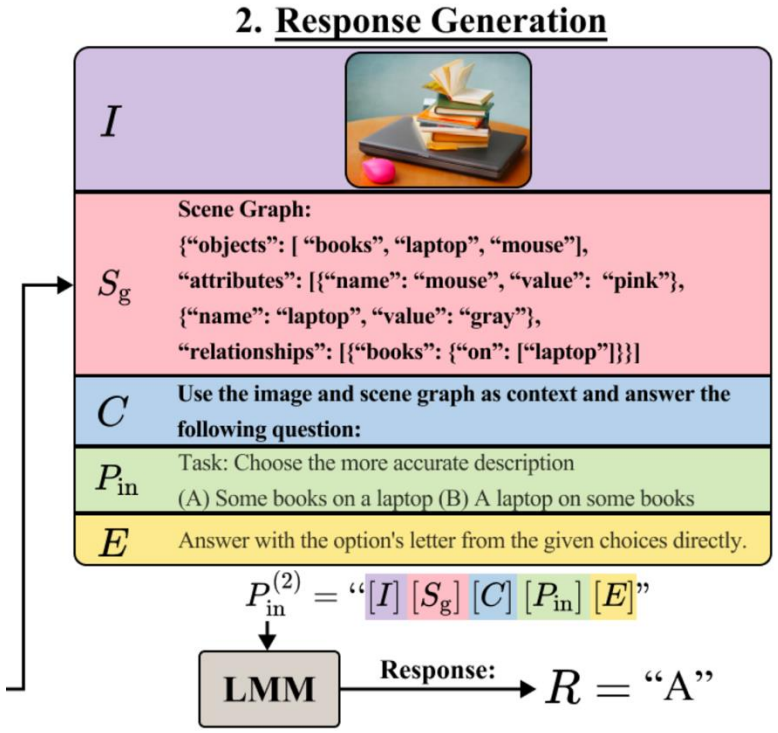
- Prompt1：$P_{\text{in}}^{(1)} = \text{“}[I]\ [P_{\text{in}}]\ [S_{\text{in}}]\text{”}$
  - I:图像
  - P_{in}：task prompt
  - S_{in}：SGG prompt
- 用task prompt限定SGG范围
- JSON格式的序列容易被LLM理解和生成



生成回答

- Prompt2： $P_{\text{in}}^{(2)} = \text{“}[I]\ [S_{\text{g}}]\ [C]\ [P_{\text{in}}]\ [E]\text{”}$

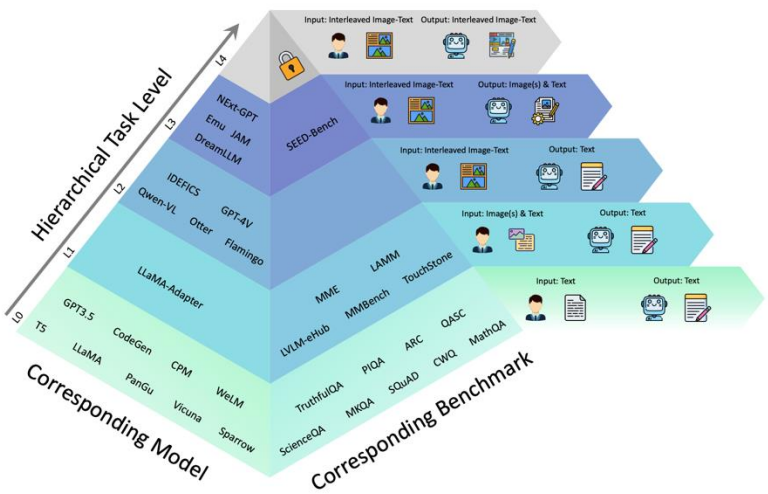
  - I:图像
  - S_g： $S_{\text{g}} = f(v_\phi(I), l(P_{\text{in}}^{(1)}))$
  - C： context prompt
  - P_{in}： task prompt
  - R： $R = f(v_\phi(I), l(P_{\text{in}}^{(2)}))$

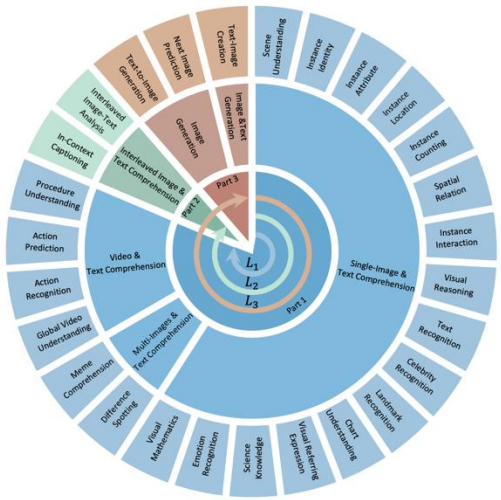


- 研究背景
- 研究方法
- 实验效果
- 后续工作
- 总结

实验设置（SEED-Bench）

- Multi-choice question
  - 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], etc.





As shown in the picture, this is the menu of McDonald's in St. Petersburg, Russia.

If I want to buy two of the burgers this girl is eating at McDonald's in St. Petersburg, how much would it cost me?

A. 130 rubles

B. 260 rubles

C. 75 rubles

D. 520 rubles

实验性能

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.

消融实验

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.

实验性能

- MM-Bench

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].

实验性能

- SEED-Bench

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]].

例子

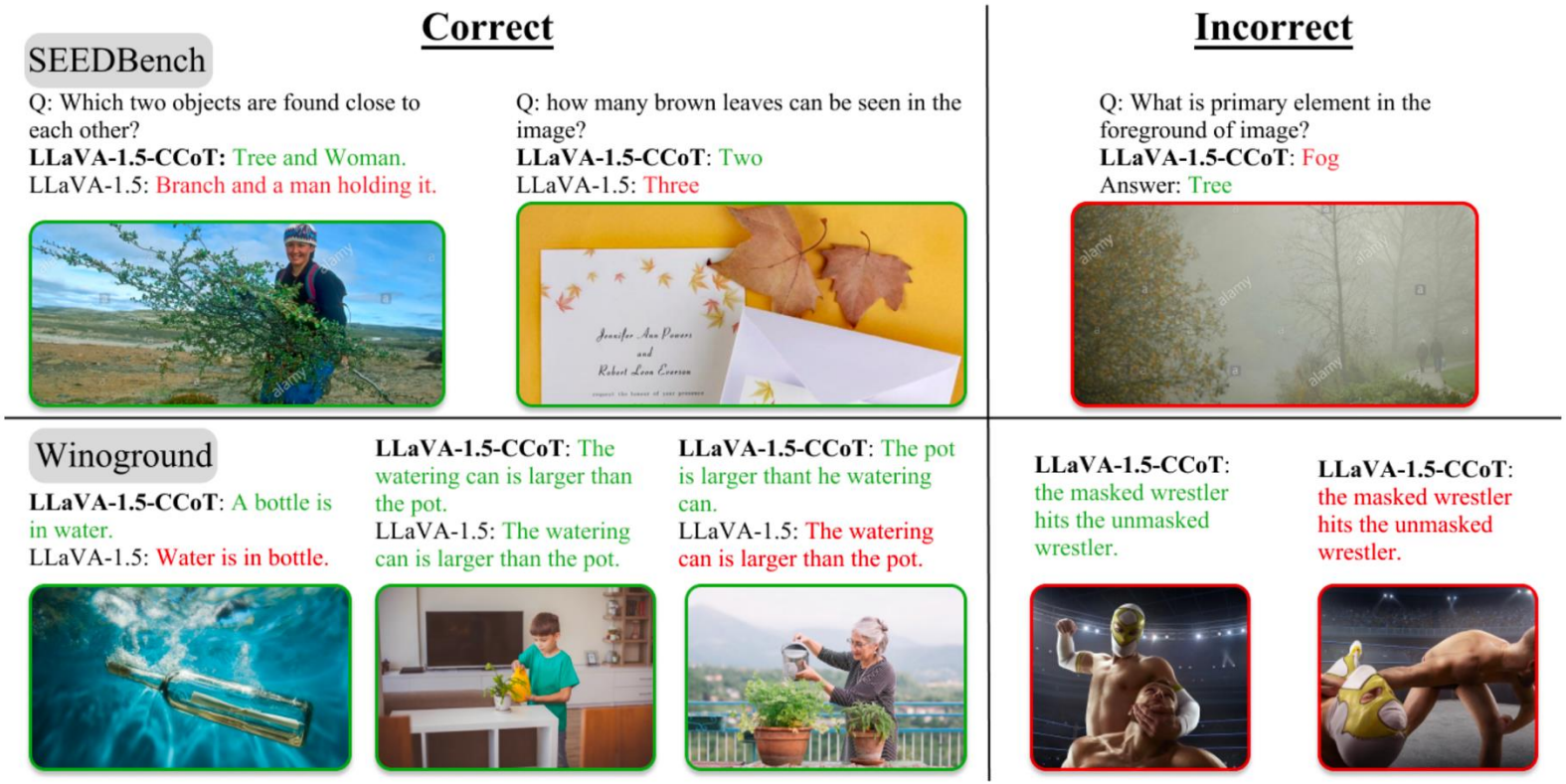


Figure 3. **Example Outputs.** Above we show examples of our method on both SEEDBench and Winoground. On the left we show successes of CCoT while the right shows failure cases. For more qualitative visualizations, please refer to Section C in Supplementary.

- 研究背景
- 研究方法
- 实验效果
- 总结

总结

- 利用LLM的开放词汇、序列处理的优势生成场景图

- 如何更好的向MLLM中注入结构化信息（例如scene graph）
  - 预训练：标注和训练成本高，且下游任务需求多样
  - 微调：　引入新的数据形式，导致灾难性遗忘问题
  - CoT：　挖掘模型的固有能力，但是没有本质上提升模型感知能力

Thanks!