Zero-Shot Vision Language Reasoning via Dual-layer Scene Graph Chain of Thoughts (Student Abstract)

Yash Bansal*, **Parshiv Kapoor***, **Agam Pandey**

Indian Institute of Technology, Roorkee, Uttarakhand, India 247667
{yash_b, parshiv_k}@bt.iitr.ac.in, agam_p@ce.iitr.ac.in

Abstract

Large Multimodal Models (LMMs) often hallucinate objects and struggle with compositional reasoning in complex visual scenes. Structured Scene Graph (SG) representations explicitly encoding objects, attributes, and relations can mitigate these issues; however, finetuning risks catastrophic forgetting. Recent zero-shot approaches prompt LMMs with scene graphs, yet typically rely on a single SG generated in one step, limiting capture of holistic context and question-specific details. We introduce a Dual-Layer Scene Graph Chain of Thoughts (DLSG-CoT) framework that enriches reasoning by combining two structured SGs: a Global Scene Graph (G-SG) that offers comprehensive image context, and a Query-Specific Scene Graph (Q-SG) produced through a two-step process targeting information relevant to the input query. Extensive experiments demonstrate that DLSG-CoT substantially improves LMM performance on compositional and context-sensitive tasks.

Introduction

Large Multimodal Models (LMMs) have shown impressive capabilities in tasks combining vision and language, yet they continue to struggle on reasoning benchmarks that require understanding. Studies have revealed that state-of-the-art models often treat images as a "bag of objects," (Ma et al. 2023), failing to capture structured interactions and dependencies between entities. Scene graphs (SGs) provide a structured representation of images, capturing the objects, their attributes, and relations (Johnson, Gupta, and Fei-Fei 2018). Several methods have explored fine-tuning on scene graphs to enhance visual reasoning in VLMs and LMMs; however, they are prone to catastrophic forgetting and loss of generalizability(Herzig et al. 2023). (Bitton-Guetta et al. 2023) has shown that scene graphs can serve as structured intermediates for zero-shot reasoning in LMMs, enabling compositional visual understanding without fine-tuning. However, it can suffer from hallucinating question-specific objects and lacking overall image context, which reduces reasoning quality and accuracy in visual question answering. We hypothesize that reasoning can be further improved by incorporating Dual-layer Scene Graphs. Specifically, a Global Scene Graph (G-SG) captures the holistic context of the image, while a Query-Specific Scene Graph (Q-SG) made through two steps extracts objects, attributes, and relationships relevant to the input query. Our method leverages these JSON-formatted and structured dual-layer scene graphs to facilitate reasoning using zero-shot prompting, termed **DLSG-CoT** (Dual Layer Scene Graph Chain of Thoughts).

*These authors contributed equally.

Methodology

For each image-query pair (I, q), we adopt a dual-layer scene graph approach (Fig. 1): a GSG captures the overall context of the image, while a QSG encodes entities and relations relevant to the query. During inference, the LMM processes the image, query, and both graphs in a structured format for response generation.

Global Scene Graph (G-SG) construction

The global scene graph (G-SG) is generated through a single zero-shot prompting pass to the LMM using only the image I and a structured prompt (P_g) that requests a comprehensive list of objects, attributes, and their relations with confidence scores in JSON format.

Query-Specific Scene Graph (Q-SG) construction

For each image-query pair (I, q), we generate a query-specific scene graph, QSG, through a two-step process. First, we extract a structured object list $\mathcal{O} = (o_i, a_i, r_i)_{i=1}^{N}$ in JSON format using prompt (P_o), where each object o_i includes its attribute a_i and relations r_i. Second, using $\mathcal{O}$, q and I as inputs, we use a structured prompt (P_q) for the LMM to refine and filter this list into a focused graph $QSG = g(\mathcal{O}, q)$ in a fixed JSON format that highlights only the objects and relations relevant to answering the query. Thus, following a dual-step scene graph generation method for query-specific scene graphs.

Response Generation

We utilize both the scene graphs, QSG and GSG, in a structured JSON format as structured intermediate representations to guide response generation. The LMM is prompted with the input image I, GSG, QSG, and the query q, using a carefully designed prompt P_{in} that encourages the

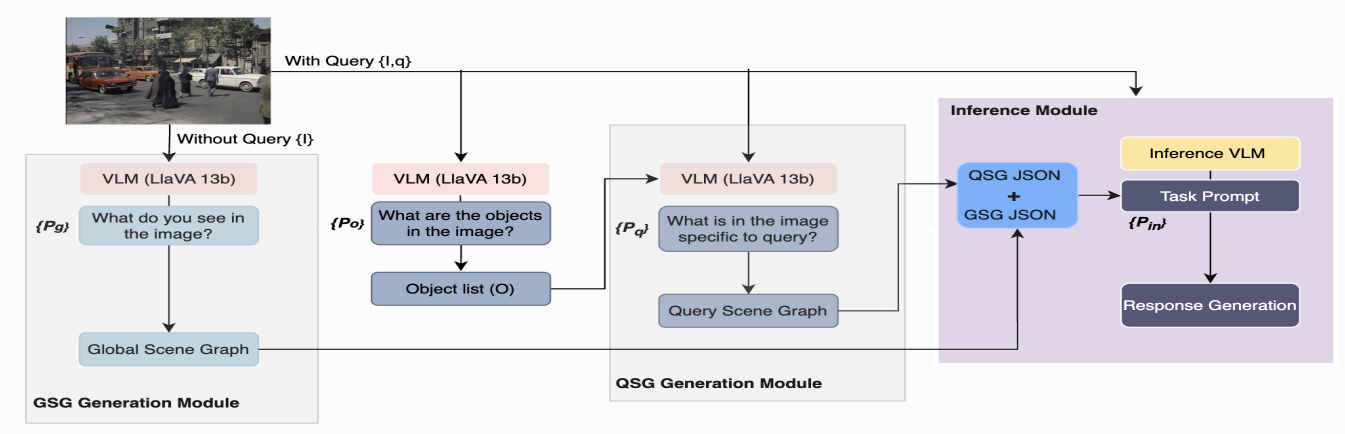

Figure 1: DLSG-CoT fuses Global and Query-Specific Scene Graphs to provide query-aware and global context for inference.

model to generate a response over the holistic image context from GSG and the query-relevant information from QSG. This dual-graph prompting approach helps the model better understand the image's overall structure and focus on the details necessary to answer the specific query. The input prompt is formally defined as:

$$P_{\text{in}} = \text{"}[I]\ [GSG]\ [QSG]\ [q]\ [N]\text{"} \quad (1)$$

where N is the instruction sentence for response generation.

Experiments and Results

Experimental Setup: We evaluate under zero-shot (ZS) settings using LLaVA-13B as a fixed scene-graph generator, with ablations on WHOOPS, SEED-Image (1k-sample subset), and MMBench (1k-sample subset) due to computational limits. We use LLaVA-1.5-13B and 7B for ablations; although our framework is model-agnostic, experiments are restricted to these models because generating scene graphs at scale is costly.

Experimental Results: Dual-Layer Scene Graphs Chain of Thoughts (DLSG-CoT) consistently outperform the single-layer CCoT scene-graph approach (Mitra et al. 2024) (Table 1). Ablations with QSG only, QSG+GSG, and no scene graph show that combining QSG and GSG improves performance in some cases but can degrade it in others relative to QSG alone. This indicates that adding GSG can sometimes dilute attention to query-specific elements, reducing reasoning accuracy. However, it is seen that QSG generated through our dual-step structured method outperforms the single-step generation approach employed by the existing method (Mitra et al. 2024).

Model	Setting:Zero-Shot	Whoops!	MM-Bench	SEED
LLaVA-1.5-13B	(w/o SG)	59.18	62.4	51.4
	(w/ SG) CCoT	60.45	63	55.8
	(w/ GSG+QSG)	64.52	**64.4**	**58**
	(w/ QSG)	**66.08**	**64.4**	56
LLaVA-1.5-7B	(w/o SG)	57.74	47.10	42.80
	(w/ SG) CCoT	59.10	52.1	49.10
	(w/ GSG+QSG)	60.47	**54.40**	**50.50**
	(w/ QSG)	**63.93**	52.70	**50.50**

Table 1: ZeroShot of LLaVA-1.5-13B and LLaVA-1.5-7B across datasets.

Future Work

We plan to extend our approach to additional compositional multimodal benchmarks, including Winoground-style datasets, to more rigorously assess generalization. We also aim to integrate reinforcement learning methods such as Direct Preference Optimization (DPO), etc., leveraging scene graph representations to strengthen reasoning. These directions will address current limitations and further advance multimodal reasoning.

References

Bitton-Guetta, N.; Shvarts, S.; Schwartz, I.; and Wolf, L. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Herzig, R.; Mendelson, A.; Karlinsky, L.; Arbelle, A.; Feris, R.; Darrell, T.; and Globerson, A. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation from Scene Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10910–10921.

Mitra, A.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.