Generating Open-World & Multi-Hierarchy Scene Graphs for Human-Instructed Manipulation Tasks via Foundation Models

Sandeep S. Zachariah*, Aman Tambi*, Moksh Malhotra, P. V. M. Rao and Rohan Paul
Indian Institute of Technology Delhi

ICRA2024 YOKOHAMA | JAPAN

Motivation and Challenges

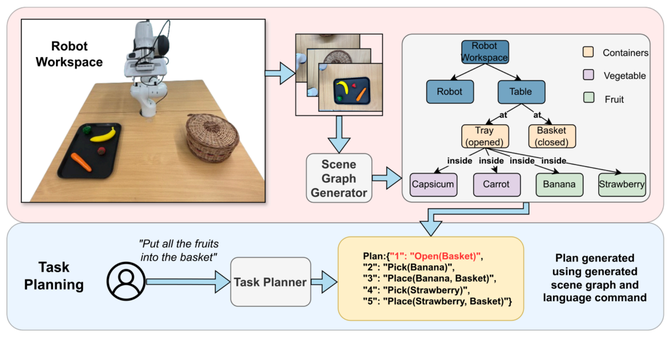
Consider a robot asked to "*bring me all the fruits*" or "*put the book which has spectacles on top of it on the rack*". Following such instructions requires the robot to possess:

- grounded semantic understanding of its environment in terms of which objects are present,
- metric location and extent of objects and,
- how objects are related to each other (inside, left of, behind, supported by etc.).

Method	Objects		Relations		
	Detection	Pose	Spatio-depth	Planar	Abstraction
GPT-4V	✓	✗	✓	✗	✓
CogVLM	✓	✓	✗	✗	✗
GPT-4V+GDINO	✓	✓	✓	✓	✓
GPT-4V+CogVLM	✓	✓	✓	✓	✓
ConceptGraph	✓	✓	✓	✗	✗
Proposed	✓	✓	✓	✓	✓

Approaches like ***Hydra***[1] use supervised learning methods to infer metric-semantic graph representations for a scene - fails for open-world settings.

Concept-Graphs[2] generates scene graphs for open-world settings but is not amenable to rapid update necessary for task execution in dynamic scenes.



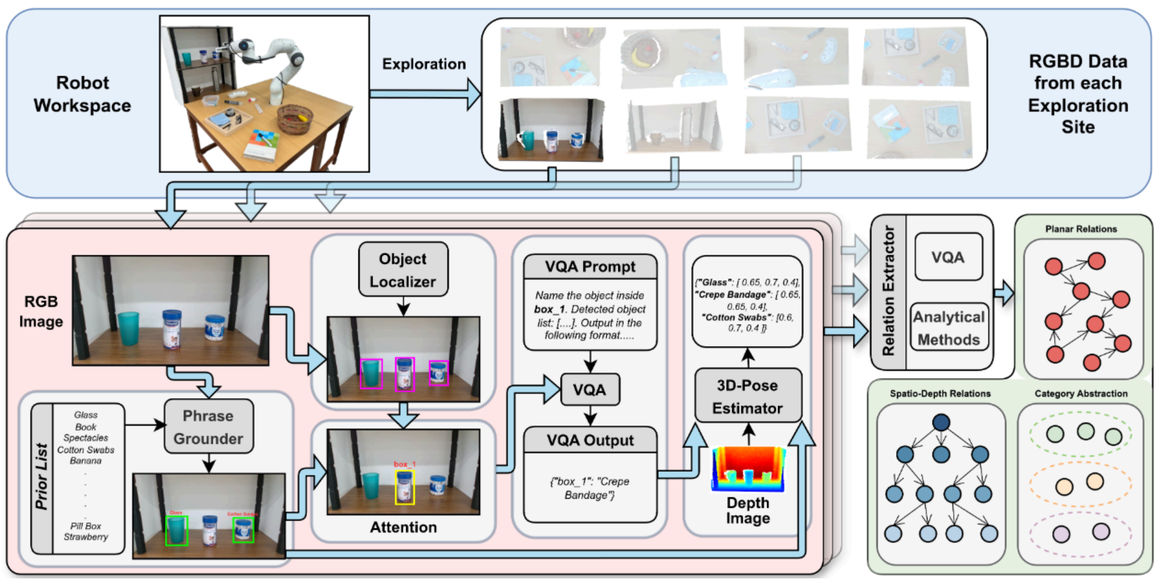
Need of a scene graph for abstract instruction following.

Technical Approach

We employ a factored approach for graph generation as (i) detecting presence and the type of objects and (ii) estimating inter-object relations.

$$\text{ObjectDetection}(\mathcal{I}_{o:T}) \rightarrow \mathcal{O} \qquad \text{RelationEstimation}(\mathcal{I}_{o:T}|\mathcal{O}) \rightarrow \mathcal{R}$$

Scene Graph Generation Pipeline



Problem Setup

The robot is operating in an environment populated by **a-priori** set of objects $o \in O$. The robot's overall task is to synthesize the sequence of actions $\pi = [a_0, a_1, .. a_n]$ in response to the natural language command $\lambda \in \Lambda$.

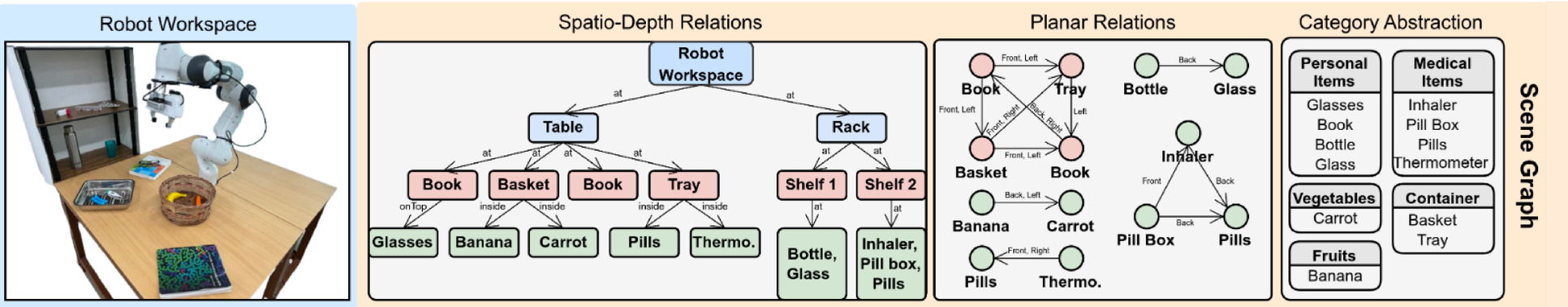
In order to generate π the robot must infer a scene graph G_t that models objects O and relations R at time t from a sequence of RGB-D information $I_{o:T}$ captured by the robot.

$$\text{SceneGraphGeneration}(\mathcal{I}_{o:T}) \rightarrow \mathcal{G}_T = (\mathcal{O}, \mathcal{R})$$

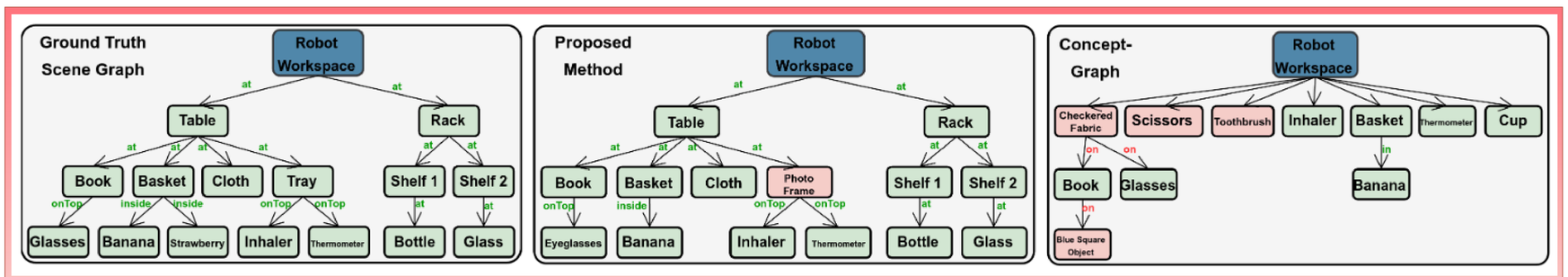
Modeling Objects : A combination of a **grounding model and a Visual Question Answering (VQA) model** is employed to detect objects. Attention is provided to these VQA models in the form of unlabelled bounding boxes, masks, or cropped images.

Modeling Relations: Spatio-depth relations ("onTop", "inside", "at"), category abstractions ("fruits", "medicinal items"), planar relations ("left", "front").
For extracting spatio-depth relations and category abstractions, a VQA model (GPT-4V) is used and planar relations are calculated analytically.

Results



Multi-hierarchical Scene Graphs: The figure shows the scene and the corresponding multi-hierarchical scene graph that was generated for the given robot workspace.

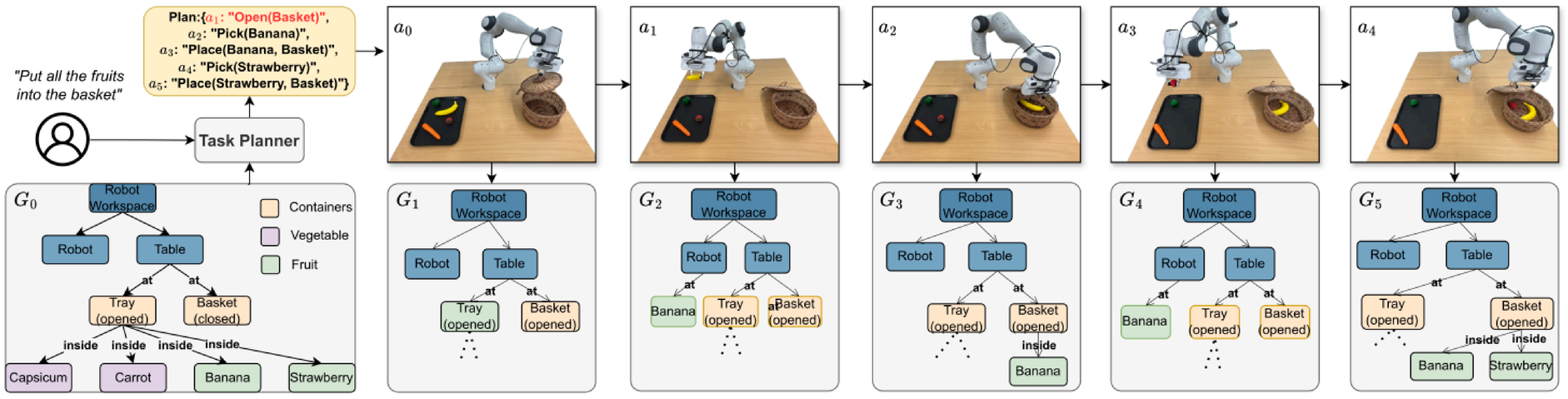


Qualitative Comparison of Scene Graphs: Proposed Method vs Concept-Graphs. The proposed method generates a scene graph closer to the ground truth. Results indicate that reliable relation extraction is dependent on object detection accuracy.

Method	Precision (%)	Recall (%)	F-measure
GDino+GPT-4V	85.0	57.0	0.68
CogVLM+GPT-4V	62.1	93.8	0.75
CogVLM	79.5	77.8	0.79
ConceptGraphs	44.8	73.2	0.54
ConceptGraphs-D	55.8	38.7	0.46
Ours	**94**	**96**	**0.95**

Accuracy of object detection. The proposed object detection pipeline has the highest F-measure showing its robustness in open-world settings

Plan Rollout: The figure shows the execution of the plan by the Franka Emika arm in response to the language instruction. The proposed scene graph supports **abstract language instructions**. It also illustrates how the scene graph gets **locally updated** after each action execution.



Conclusion

We introduce a novel approach for generating multi- hierarchical scene graphs for **open-world** settings which facilitates the execution of **sequential and long-horizon plans**. It employs a **factored approach** to scene graph generation, first detecting objects and then establishing relations between them.

Acknowledgement

Authors acknowledge research support from from the National Center for Assistive Health Technologies (NCAHT), IIT Delhi supported by the Indian Council for Medical Research (ICMR), Govt. of India and the DRDO-IIT Delhi Defence Industry Academia Center for Excellence (DIA-COE), Govt. of India.

References

[1] N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. arXiv preprint arXiv:2201.13360, 2022.
[2] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen,A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. arXiv preprint arXiv:2309.16650, 2023.

Website