# ASHiTA: Automatic Scene-grounded HIerarchical Task Analysis

Yun Chang[1], Leonor Fermoselle[2], Duy Ta[2], Bernadette Bucher[3], Luca Carlone[1], Jiuguang Wang[2]
[1]Massachusetts Institute of Technology, [2]Robotics and AI Institute, [3]University of Michigan

Paper

## Research Question:

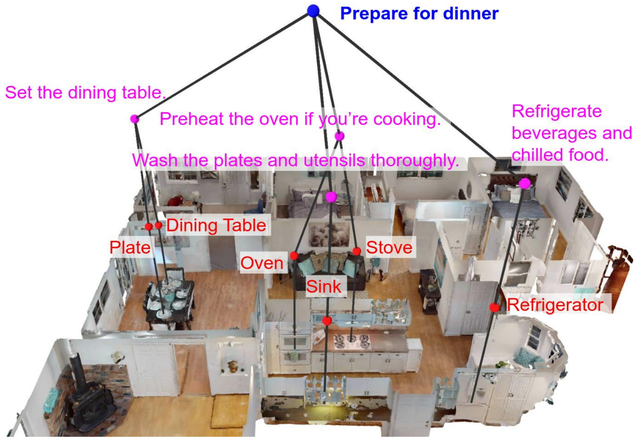
How can we enable robots to understand **complex high-level commands** and **ground** them in **real-world 3D scenes**?

## Contributions:

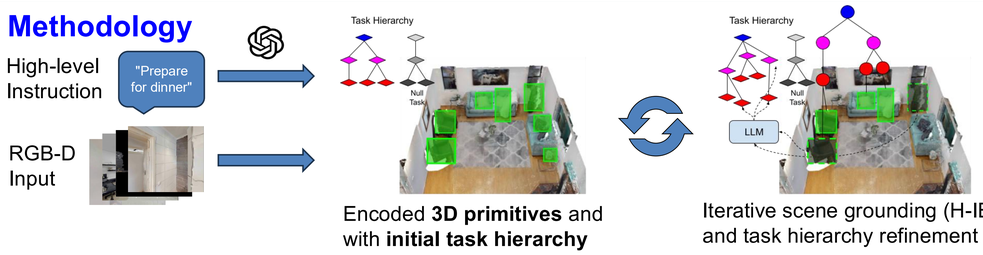
- **Generalize the Information Bottleneck (IB) principle** to build a **hierarchical, task-defined 3D scene graph** based on a given task decomposition.
- **Iterative framework** alternating LLM-based **task decomposition** and **scene graph construction** to generate a scene grounded task hierarchy.

## Key Results:

- **State-of-the-art grounding accuracy** of task hierarchies to task-relevant objects, **outperforming existing zero-shot models**.
- **Outperforms** LLM + open-set scene graph baselines in **scene-grounded task hierarchy generation**.




## Methodology




## Hierarchical Information Bottleneck (H-IB):

$$\min_{\mathbb{P}(\mathcal{S}|\mathcal{S}_0)} I(\mathcal{S}_0;\mathcal{S}) - \beta I(\mathcal{T};\mathcal{S})$$

Generalization to *n* **hierarchical layers**

$$\min_{\mathbb{P}(\mathcal{S}_k|\mathcal{S}_{k-1}),\, k=1\ldots n} \sum_{k=1}^{n} I(\mathcal{S}_{k-1};\mathcal{S}_k) - \beta \sum_{k=1}^{n} I(\mathcal{T}_k;\mathcal{S}_k)$$

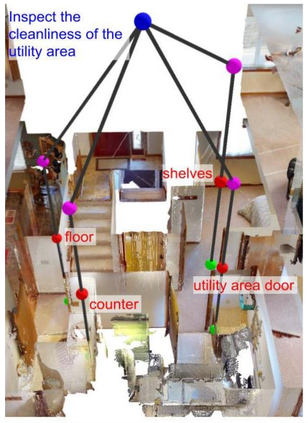
Iterative multi-layer update steps

$$\begin{cases} p_\tau(s_k|s_{k-1}) = \frac{1}{Z} p_\tau(s_k)\exp(-\beta d) \\ p_{\tau+1}(s_k) = \sum_{s_{k-1}} p_\tau(s_{k-1}) p_\tau(s_k|s_{k-1}) \\ p_{\tau+1}(t_k|s_k) = \sum_{s_{k-1}} p_\tau(t_k|s_{k-1}) p_\tau(s_{k-1}|s_k) \end{cases}$$

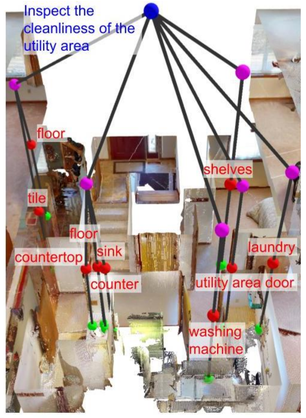
$$d = D_{KL}(p_\tau(t_k|s_k)||p_\tau(t_k|s_{k-1})) + \sum_{i=k+1}^{n} \sum_{s_i} p_\tau(s_i|s_k) D_{KL}(p_\tau(t_i|s_i)||p_\tau(t_i|s_{k-1}))$$

## Iterative Refinement:

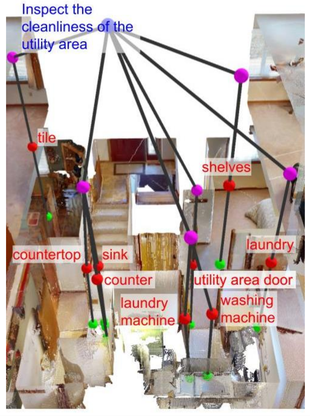
Open the utility area door: [utility area door]
Check the floor for debris: [floor]
Examine shelves for dust accumulation: [shelves]
Inspect counter for spills: [counter]
Look at bins for overflow: [bins]




Open the utility area door: [utility area door]
Check the floor for debris: [floor, tile]
Examine shelves for dust accumulation: [shelves]
Inspect counter for spills: [counter, sink, countertop]
Check for dirty laundry: [laundry]
Inspect the washing machine:[washing machine]




Open the utility area door: [utility area door]
Check the floor for debris: [tile]
Examine shelves for dust accumulation: [shelves]
Inspect counter for spills: [counter, sink, countertop]
Check for dirty laundry: [laundry]
Inspect the washing machine:[washing machine, laundry machine]




## Experiments:

| Method | s-acc (%) | t-acc (%) |
|---|---|---|
| 3D-VisTA [52] | 25.3 | 10.3 |
| PQ3D [53] | 24.4 | 9.7 |
| ASHiTA | **28.71** | **12.13** |
| ASHiTA + Txt Emb. | 65.4 | 39.33 |
| GPT w/ GT labels [50] | **75.9** | **52.1** |

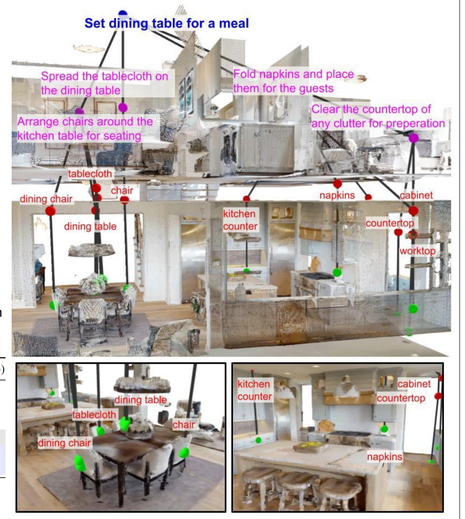
Table 1. Evaluation of grounding using the SG3D HM3DSem dataset [1] with ground truth instances against SOTA zero-shot models [2, 3].

| Method | s-acc (%) | t-acc (%) |
|---|---|---|
| Hydra [14] + GPT | 8.18 | 2.44 |
| HOV-SG [45] | 8.98 | 1.95 |
| ASHiTA | **21.7** | **8.78** |
| Hydra (GT Seg) + GPT | 14.2 | 6.34 |

Table 2. Evaluation of grounding on 8 scenes from the SG3D HM3DSem dataset [1] with RGB-D Input against LLM + Scene Graph baselines [4, 5]

| Method | s-rec (%) | s-prec (%) | t-acc (%) |
|---|---|---|---|
| Hydra + GPT | 9.43 | 15.51 | 4.88 |
| HOV-SG + GPT | 4.55 | 4.87 | 1.95 |
| ASHiTA | **10.39** | **20.6** | **9.27** |
| ASHiTA (GT Pos) | 15.12 | 34.47 | 16.59 |
| Hydra (GT Seg) [14] + GPT | 17.06 | 18.98 | 14.63 |
| ASHiTA (GT Pos + Txt Emb) | **38.71** | **34.39** | **36.1** |

Table 3. Evaluation of Hierarchical Task Analysis. Against LLM + Scene Graph baselines [4, 5]







## Limitations / Future Work

- Generated task hierarchies are not guaranteed to be **feasible** or **complete** to accomplish the high-level tasks
- The structure of task hierarchy prevents objects from being shared across subtasks, reducing flexibility and descriptiveness.
- **Future direction**: 1) Integrate classical planning techniques to validate and refine subtask sets to ensure **task completeness**. 2) Bridge the task hierarchy generation with **robot skills**.

## References

[1] Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tengyu Liu, Xiao-jian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding in 3d scenes. ArXiv preprint: 2408.04034, 2024.
[2] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 2899–2909, 2023.
[3] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. arXiv preprint: 2405.11442, 2024.
[4] N. Hughes, Y. Chang, and L. Carlone. Hydra: a real-time spatial perception engine for 3D scene graph construction and optimization. In Robotics: Science and Systems (RSS), 2022.
[5] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. Robotics: Science and Systems (RSS), 2024.