S3E: Semantic Symbolic State Estimation With Vision-Language Foundation Models

Guy Azran[1,2], Yuval Goshen[1], Kai Yuan[2], Sarah keren[1]
[1]Taub Faculty of Computer Science, Technion – Israel Institute of Technology
[2]Extended Realities Lab, Intel Labs
{guy.azran, yuval.goshen}@campus.technion.ac.il, kai.yuan@intel.com, sarahk@cs.technion.ac.il

***Abstract*—In automated task planning, symbolic state estimation is the process of translating sensor input into a high-level task state. This is especially important in robotic settings where unpredictable environments and actions often lead to unexpected outcomes. State estimation enables the agent to manage uncertainties, adjust its plans, and make more informed decisions. Traditionally, researchers and practitioners relied on handcrafted and hard-coded state estimation functions to determine the abstract state defined in a specific task domain. Advancements in Vision-Language Models (VLMs) enable autonomous retrieval of semantic information from visual input. We present Semantic Symbolic State Estimation (S3E), the first general-purpose symbolic state estimator based on VLMs that can be applied in various robot task settings without specialized coding or additional exploration. S3E takes advantage of the foundation model's internal world model and semantic understanding to assess the likelihood of certain symbolic components of the environment's state. We analyze S3E as a multi-label classifier, reveal different kinds of uncertainties that arise when using it, and show how they can be mitigated using natural language and targeted environment design. While our method is generic, we aim to facilitate symbolic state estimation in robotic settings. We show that S3E can achieve over 90% state estimation precision in our simulated and real-world robot experiments.**

I. Introduction

Automated task planning is crucial for intelligent robotic agents to solve complex and ever-changing tasks [5, 4]. In some cases, it is assumed that an agent has full domain knowledge (the Closed World Assumption (CWA) [10]) and that all task-related facts are known. However, an agent's observations are often based on sensing capabilities from which extracting these facts is non-trivial. This is especially true in real-world robot applications. Symbolic state estimation is the process of obtaining a high-level state of the environment, i.e., translating numeric sensor input into semantic facts [2, 1, 6]. This helps monitor plan execution; an agent reaching an unexpected state may be grounds for replanning or reporting of task failure. This is particularly important in robotic settings where agents perform error-prone motions in the physical world.

Example 1: A robotic arm is tasked with rearranging groceries on multiple tables. The goal is to move a box of cereal and a carton of milk to a specific table where the hungry human would like to prepare her breakfast. A task planner chooses the following plan: "pick-up(milk, table1)", "put-down(milk, table3)", "pick-up(cereal, table2)", "put-down(cereal, table3)". While moving to place the cereal on table number 3, the object is dropped due to an unstable grasp and lands on table number 1. Using a state estimator, we detect an unexpected state: the cereal box is not on table number 3. We thus call the task planner once more to obtain the following plan that will lead us to the goal state: "pick-up(cereal, table1)", "put-down(cereal, table3)".

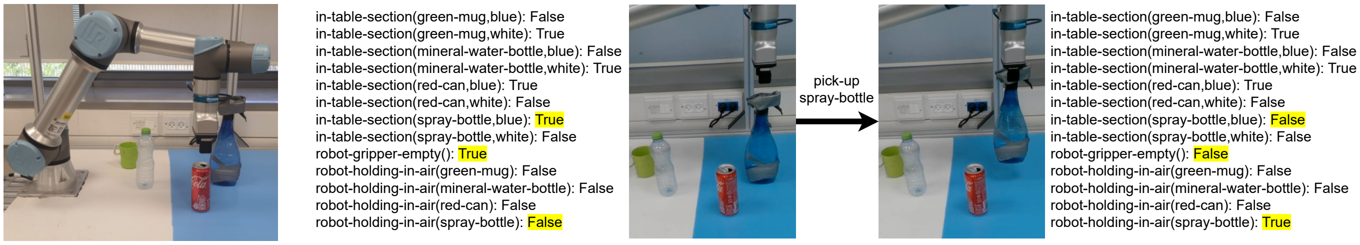
Current state-of-the-art task planning methods rely on handcrafted and hard-coded state estimation functions [8, 3]. This is time-consuming and relies on expert domain knowledge and advanced sensing equipment, which results in domain-specific outputs that do not adapt to environment or task changes. We desire a general state estimation function that requires no specialized coding, no additional exploration, and generalizes to a large scope of tasks.

With the rise of powerful instruction-based Vision-Language Models (VLMs), i.e., vision-based foundation models, it is now possible to answer complex semantic questions about a scene based on visual input alone [7, 9]. Previous approaches required a specialized combination of computer vision tools to answer specific question sets. By comparison, VLMs are designed to answer any question. Questions are specified in natural language and are mostly answered accurately if the input is within its training distribution.

We introduce Semantic Symbolic State Estimation (S3E), the first zero-shot state estimator based on VLMs. Our objective is to provide a general, versatile, and performant solution for state estimation that will accelerate the construction of state estimation functions for researchers and practitioners of task planning. S3E exploits the foundation model's internal world model and semantic understanding [12, 13] to assess the likelihood of task-related symbolic components of the environment's state as the agent manipulates the physical world. It consists of two stages: (1) translating symbolic predicate definitions into natural language questions and (2) answering them given visual input. We show that the translation stage significantly improves performance in Appendix **??**.

Fig. 1 demonstrates the usage of S3E in a real-world robotics task. Our experiments show that a high-accuracy zero-shot state estimation solution is possible. While this approach is highly generic, we provide a video demonstration of how S3E can be deployed in robotic manipulation tasks, available as supplementary material[1]. To improve performance

[1] In the supplementary video we demonstrate how S3E is used for plan execution monitoring, action failure detection, and task failure detection


in-table-section(green-mug,blue): False
in-table-section(green-mug,white): True
in-table-section(mineral-water-bottle,blue): False
in-table-section(mineral-water-bottle,white): True
in-table-section(red-can,blue): True
in-table-section(red-can,white): False
in-table-section(spray-bottle,blue): True
in-table-section(spray-bottle,white): False
robot-gripper-empty(): True
robot-holding-in-air(green-mug): False
robot-holding-in-air(mineral-water-bottle): False
robot-holding-in-air(red-can): False
robot-holding-in-air(spray-bottle): False
pick-up
spray-bottle
in-table-section(green-mug,blue): False
in-table-section(green-mug,white): True
in-table-section(mineral-water-bottle,blue): False
in-table-section(mineral-water-bottle,white): True
in-table-section(red-can,blue): True
in-table-section(red-can,white): False
in-table-section(spray-bottle,blue): False
in-table-section(spray-bottle,white): False
robot-gripper-empty(): False
robot-holding-in-air(green-mug): False
robot-holding-in-air(mineral-water-bottle): False
robot-holding-in-air(red-can): False
robot-holding-in-air(spray-bottle): True

(a) Setup.

(b) Example transition annotated with S3E. State changes highlighted.

Fig. 1. Visual results from a robotic pick-and-place task using S3E - after picking up the spray bottle, 'robot-gripper-empty()' and in-table-section(spray-bottle, blue)' are set from True to False. We refer the reader to the supplementary materials for a demo video of this example.

alongside the VLM's strong priors, we use natural language instruction and targeted environment design to remove ambiguities and reduce uncertainties about the environment or task.

We analyze S3E as a multi-label classifier where the labels are the set of grounded predicates that make up all possible facts about the task state. Our experiments focus on usability and showcase a simulated and real-world robotic domain. S3E also achieves high performance in a photorealistic blocksworld with ever-changing objects in the Appendix, showing that it is truly general-purpose and versatile. We propose task-specific solutions to handle two kinds of uncertainties in our proposed state estimator. The first is the model's uncertainty regarding the state. The second stems from the subjective nature of the actual state relative to the intent of the task designer, i.e., whether a certain property holds for a given state is in the eye of the beholder. We show examples of these uncertainties and how they can be reduced using natural language instruction and minimal environment design. This improves on previous work that elicit uncertainties in language models [11, 14] by leveraging this idea for symbolic state estimation in the context of task planning. Regardless of these uncertainties, general-purpose state estimation is a needed change from the specialized solutions offered by today's state-of-the-art.

This paper presents the following contributions:

- Introduction of Semantic Symbolic State Estimation (S3E): first zero-shot symbolic state estimator using vision-based foundation models.
- Proposal of a general solution for high-level state estimation in task planning.
- Identification and mitigation of model uncertainty and task-specific ambiguity.
- Empirical demonstration of S3E's effectiveness in simulated and real-world environments.

A video demo of S3E is available at the following link: https://drive.google.com/file/d/1meD4Yg3l1gNjmwjP-C9-dYPSfVeeTf2R/view?usp=sharing

References

[1] Nicola Castaman, Enrico Pagello, Emanuele Menegatti, and Alberto Pretto. Receding Horizon Task and Motion Planning in Changing Environments. *Robotics and Autonomous Systems*, 145:103863, November 2021. ISSN 0921-8890. doi: 10.1016/j.robot.2021.103863.

[2] Siwei Chen, Anxing Xiao, and David Hsu. LLM-State: Open World State Representation for Long-horizon Task Planning with Large Language Model, April 2024.

[3] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. PDDLStream: Integrating Symbolic Planners and Blackbox Samplers via Optimistic Adaptive Planning, March 2020.

[4] Hector Geffner and Blai Bonet. *A Concise Introduction to Models and Methods for Automated Planning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2013. ISBN 978-3-031-00436-0 978-3-031-01564-9. doi: 10.1007/978-3-031-01564-9.

[5] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated Planning and Acting*. Cambridge University Press, Cambridge, 2016. ISBN 978-1-107-03727-4. doi: 10.1017/CBO9781139583923.

[6] Fabien Lagriffoul, Neil T. Dantam, Caelan Garrett, Aliakbar Akbari, Siddharth Srivastava, and Lydia E. Kavraki. Platform-Independent Benchmarks for Task and Motion Planning. *IEEE Robotics and Automation Letters*, 3(4):3765–3772, October 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2856701.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023.

[8] Magí Dalmau Moreno, Néstor García, Vicenç Gómez, and Héctor Geffner. Combined Task and Motion Planning via Sketch Decompositions. *Proceedings of the International Conference on Automated Planning and Scheduling*, 34:123–132, May 2024. ISSN 2334-0843. doi: 10.1609/icaps.v34i1.31468.

[9] OpenAI. GPT-4V(ision) system card. https://openai.com/index/gpt-4v-system-card/, 2023.

[10] Raymond Reiter. ON CLOSED WORLD DATA BASES. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 119–140. Morgan Kaufmann, January 1981. ISBN 978-0-934613-03-3. doi: 10.1016/B978-0-934613-03-3.50014-3.

[11] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. In *Proceedings of The 7th Conference on Robot Learning*, pages 661–682. PMLR, December 2023.

[12] Johannes Schneider, Christian Meske, and Pauline Kuss. Foundation Models. *Business & Information Systems Engineering*, 66(2):221–231, April 2024. ISSN 1867-0202. doi: 10.1007/s12599-024-00851-0.

[13] Alan F. Smeaton. Understanding Foundation Models: Are We Back in 1924?, September 2024.

[14] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, October 2023.