LIT: Large Language Model Driven Intention Tracking for Proactive Human-Robot Collaboration - A Robot Sous-Chef Application

Zhe Huang John Pohovey Ananya Yammanuru Katherine Driggs-Campbell
University of Illinois at Urbana-Champaign*
{zheh4, jpohov2, ananyay2, krdc}@illinois.edu

Abstract

Large Language Models (LLM) and Vision Language Models (VLM) enable robots to ground natural language prompts into control actions to achieve tasks in an open world. However, when applied to a long-horizon collaborative task, this formulation results in excessive prompting for initiating or clarifying robot actions at every task step. We propose Language-driven Intention Tracking (LIT), leveraging LLMs and VLMs to model the human user's long-term behavior and to predict the next human intention to guide the robot for proactive collaboration. We demonstrate smooth coordination between a LIT-based collaborative robot and a human in collaborative cooking tasks.

1. Introduction

The groundbreaking advances in Large Language Models (LLM) and Vision Language Models (VLM) endow robots with exceptional cognition capabilities and reasoning skills to both understand the surrounding open world and follow natural language commands of human users [2, 5]. More recent works explore conversations between the human user and the robot to allow the robot to perform multi-step tasks or clarify ambiguity of the human command [10, 12].

When the philosophy of grounding natural language commands into robot control policies is applied to human-robot collaboration (HRC), the human user may have to have a conversation with the robot at each step of the long-horizon task [12]. This situation rarely happens in human-human collaboration, as a human is able to track the progress on the partner's side based on their shared knowledge over the task. For examples, a worker rarely has to have a conversation with a co-worker in a collaborative assembly task on which they have collaborated many times, and a sous-chef rarely has to have a conversation with the chef when creating a regular dish together.

To address this challenge in human-robot collaboration, the robot needs to build an effective understanding of not only the environment, but also the human user. This work proposes Language-driven Intention Tracking (LIT) to model long-term behavior of the human user, and integrates LIT into an LLM-driven collaborative robot framework. LIT extends intention tracking [3] by applying an LLM to model measurement likelihood and transition probabilities in the probabilistic graphical model of human intentions, which is defined by grounding an overall task prompt (e.g., make a salad) with understanding of the scene using LLM and VLM models. Note this is the only prompt needed from the human user in LIT framework. LIT uses a VLM to generate text descriptions of the human user's behavior in the frames as measurements to track the human user's intention and filter out hallucinations. Intention prediction in the near-term allows the collaborative robot to proactively assist the human user. By harnessing foundation models, we believe the LIT framework can generalize to any collaborative tasks. We demonstrate the effectiveness of the LIT framework in a scenario where the collaborative robot acts as a sous-chef to assist a human user in cooking.

2. Language-driven Intention Tracking

Problem Formulation. The human user wants to make a dish, but all the required materials and tools are not reachable by the human but the robot. The robot needs to act as a sous-chef to coordinate with the human by passing essential materials and tools at appropriate times while not making the human's cooking table overly occupied with unnecessary items at the moment. The robot is assumed to only receive the prompt at the beginning on what dish is going to be made, and will not receive prompts during collaboration.

System Overview. As presented in Fig. 1, the open scene understanding module detects objects in the scene and generate potential grasp options. The task graph reasoning module takes the detected objects and the overall task prompt as input to generate a list of task steps, which we define as *intention*. As some steps of the overall task can switch order without impact on the outcome, the LLM checks on reversibility of task step sequences, and builds a task graph. The Language-driven Intention Tracking module uses the task graph to build the probabilistic graphical

*This work was supported by the National Science Foundation under Grant No. 2143435.

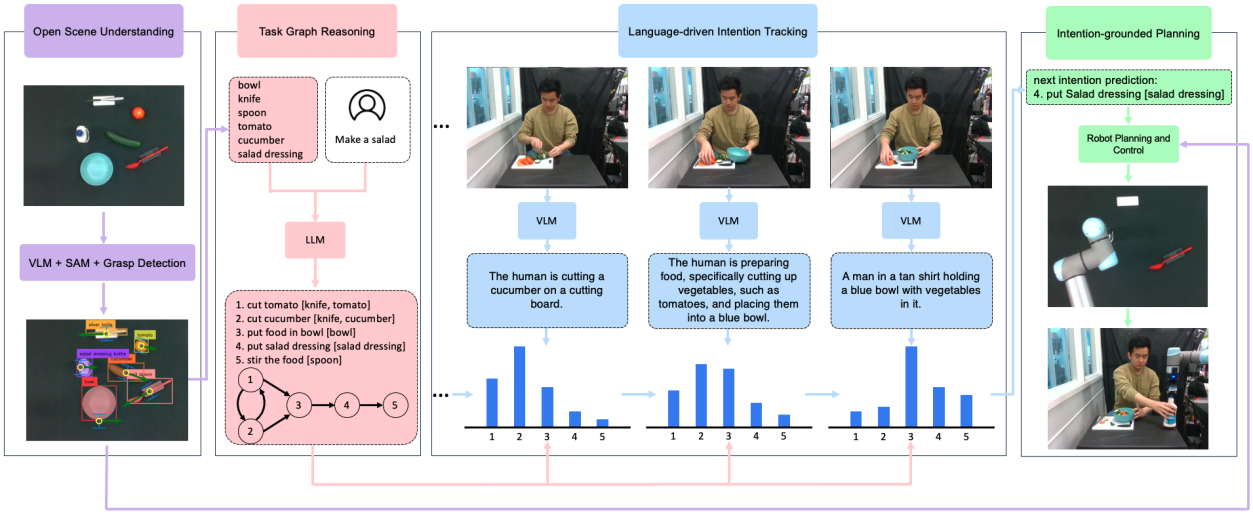

Figure 1. Language-driven Intention Tracking (LIT) based collaborative robot framework.

model for intention transition. The VLM is used to generate text descriptions from frames as measurements. We compute time-varying transition probabilities and make prediction steps, and use measurements to compute measurement likelihood and make update steps to track the human intention. The intention-grounded planning module makes an additional prediction step on the current intention posterior, and manipulate the objects relevant to the predicted next intention to proactively collaborate with the human.

We choose LLaVA [6–8] with a 13-billion parameter Vicuna [1] backbone (derived from Llama 2 [11]) as the VLM in the system. Note that we use the same model as the LLM for consistent performance by inputting the text prompt with a full-black image. The VLM is cascaded with Grounding DINO [9], Segment Anything (SAM) [4], and principal component analysis to detect, segment all objects and perform grasp synthesis. The collaborative robot is a UR5e arm equipped with a Robotiq Hand-E Gripper. We use Intel RealSense RGBD cameras to provide a top-down view of the robot table with objects on it, and to provide a front view of the human user's behavior. We use Robot Operating System (ROS) to build the LIT framework.

Language-driven Intention Tracking Method. We define human intention at time t as a discrete variable G_t. During intention tracking, we iterate prediction and update steps of Bayesian filtering as presented in Eq. 1 to get the posterior of the human intention G_t conditioned on measurements $X_{1:t}$.

$$P(g_{t+1}|x_{1:t}) = \sum_{g_t} P(g_{t+1}|g_t, x_{1:t}) P(g_t|x_{1:t})$$
$$P(g_{t+1}|x_{1:t+1}) \propto P(x_{t+1}|x_{1:t}, g_{t+1}) P(g_{t+1}|x_{1:t}) \quad (1)$$

We introduce Language Probabilistic Graphical Model to perform intention tracking, where g_t is the text of the task step the human user is performing at time t, and x_t is the textual description of the human behavior in the camera frame at time t. To calculate a conditional probability with language random variables (e.g. $P(A = a|B = b, C = c)$), we create prompts of two part: the conditional part describes $B = b$ and $C = c$, and the query part asks about A and a. We propose two methods to compose the query part: point estimate and distribution approximation.

The point estimate method asks the LLM to generate one value of A by "what do you think A would be?", and compute a similarity score [13] of the generated text with respect to a to quantify $P(a|b, c)$. The distribution approximation method requires A to be discrete. This method asks the LLM to sample a list of values of A by "what do you think A would be? Provide N examples.", allocates these values to the closest possible values of A, and forms a statistical estimate of the $P(a|b, c)$. We use the point estimate method to calculate measurement likelihood $P(x_{t+1}|x_{1:t}, g_{t+1})$, and the distribution approximation method to calculate intention transition probability $P(g_{t+1}|g_t, x_{1:t})$. Fig. 2 presents LIT in a human cooking demonstration.

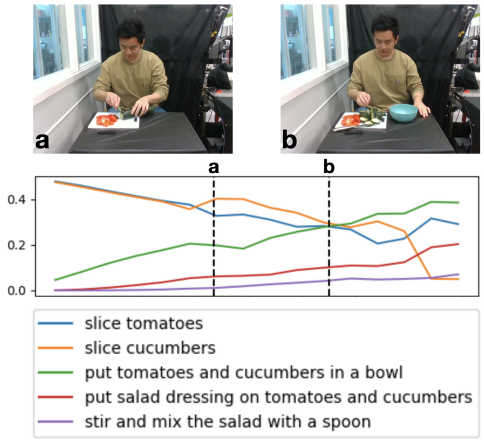

Figure 2. LIT in a human cooking demonstration. Snapshots show the moments when the human intention changes.

References

[1] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[2] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pages 540–562. PMLR, 2023. 1

[3] Zhe Huang, Ye-Ji Mun, Xiang Li, Yiqing Xie, Ninghan Zhong, Weihang Liang, Junyi Geng, Tan Chen, and Katherine Driggs-Campbell. Hierarchical intention tracking for robust human-robot collaboration in industrial assembly tasks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9821–9828, 2023. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2

[5] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023. 1

[6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[8] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023. 2

[10] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, pages 661–682. PMLR, 2023. 1

[11] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 2

[12] Huaxiaoyue Wang, Kushal Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, et al. Mosaic: A modular system for assistive and interactive cooking. *arXiv preprint arXiv:2402.18796*, 2024. 1

[13] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019. 2