AToM-Bot: Embodied Fulfillment of Unspoken Human Needs with Affective Theory of Mind

Fanhong Li*
Tsinghua University
lifanhong.ai@gmail.com

Wei Ding*
Tsinghua University
dw21@mails.tsinghua.edu.cn

Ziteng Ji
University of California Berkeley
zitengjiender0618@gmail.com

Zhengrong Xue
Tsinghua University
xzr23@mails.tsinghua.edu.cn

Jia Liu†
Tsinghua University
liujiaTHU@tsinghua.edu.cn

***Abstract*—We propose AToM-Bot, a novel task generation and execution framework for proactive robot-human interaction, which leverages the human mental and physical state inference capabilities of the Vision Language Model (VLM) prompted by the Affective Theory of Mind (AToM). Without requiring explicit commands by humans, AToM-Bot proactively generates and follows feasible tasks to improve general human well-being. When around humans, AToM-Bot first detects current human needs based on inferred human states and observations of the surrounding environment. It then generates tasks to fulfill these needs, taking into account its embodied constraints. We designed 16 daily life scenarios spanning 4 common scenes and tasked the same visual stimulus to 59 human subjects and our robot. We used the similarity between human open-ended answers and robot output, and the human satisfaction scores to metric robot performance. AToM-Bot received high human evaluations in need detection (6.42/7, 91.7%), embodied solution (6.15/7, 87.8%) and task execution (6.17/7, 88.1%). We show that AToM-Bot excels in generating and executing feasible plans to fulfill unspoken human needs. Videos and code are available at https://affective-tom-bot.github.io/.**

I. Introduction

How can robots fulfill human needs *without being directly told what to do?* Accurately modeling human needs and generating executable plans under physical constraints is a challenge in Human Robot Interaction (HRI). In this work, we studied this challenge by leveraging the vision language model and affective Theory of Mind.

Language is an interactive interface for humans to communicate with robots. There is considerable research on leveraging large language models for robot task planning [1, 2, 3, 4, 5, 6] and control grounding [7, 8, 9, 10, 11, 12, 13]. However, most of work begins with a clear verbal command directed at the robot. In real life, many human needs emerge before human formalize them into explicit verbal instructions; for example, when focusing on work, feeling like a cup of tea, we might reach for the tea cup while staring at the computer screen. When we find the tea cup is already empty, we might put it down and continue working rather than pausing work to make a tea. If there's a robot nearby, we may as well not bothering ask it to make tea. But if a robot is proactively observing human, it's not hard for it to infer human current need for "something to drink", then the robot can help satisfy the need within its ability.

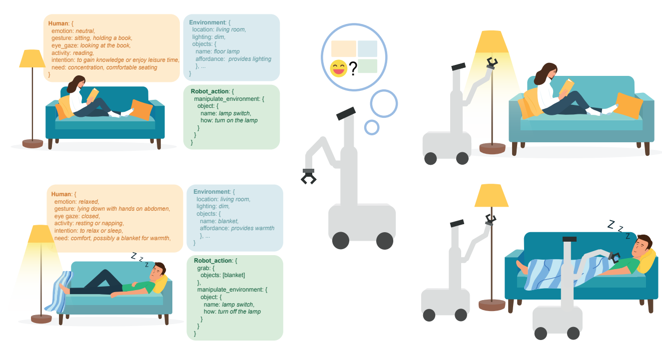

Fig. 1. AToM-Bot is a novel task generation and execution framework for proactive robot-human interaction, towards the embodied fulfillment of unspoken human needs.

There is work focusing on inferring human intention with non-verbal cues with the theory of mind [14, 15, 16, 17], but not yet grounded on a real-world robot. Such human-agent collaboration tasks require active participation in the same task. However, most of the time, people are engaged in solo tasks, focusing mainly on the task at hand, but needs beyond their main task emerge. For example, while immersive reading, one might ignore the time passing by and the darkening sky; the need for lighting emerges for humans to continue ongoing tasks. Since these tasks are relatively minor compared to the main task, people may not turn them into spoken tasks or address these spontaneous needs. Or when a person falls asleep on the sofa, they can no longer take initiative. This is especially the case for the elderly, or the disabled, who may need help the most but sometimes cannot even instruct a robot to act. At this time, a robot that understands and observes can use multimodal information to detect human needs and help humans fulfill these non-primary task needs, thus better enabling humans to perform their primary needs.

Our method proposes **AToM-Bot** - Affective Theory of Mind robot - a framework for satisfying unspoken human needs utilizing Vision Language Model (VLM) [18] prompted with Affective Theory of Mind [19, 20]. We make the following contributions: (1) Given a visual input of a human and their surrounding environment, we use VLM to detect human needs and generate embodied solutions to address these needs. (2) We grounded AToM-Bot with open-vocabulary navigation and imitation learning. (3) We designed 16 common daily scenarios to evaluate robot performance with human satisfaction rating and similarity to human response. AToM-Bot achieved 91.7%, 87.8% and 88.1% satisfaction in need detection, embodied solution and task execution. AToM-Bot represents a new potential for future coexistence and interaction between humans and robots in everyday life.

II. Related work

A. Non-verbal Communication

The deployment of LLM in robotics has empowered better translation of human commands to robot action; people can communicate with robots using text [1, 4, 5], spoken language or even yelling [21]. But apart from the explicit direct communication through language, which is highly abstract and has been through a thoughtful process, and easy to be grounded into actions, there are a significant part of human communication happened in a non-verbal way; it can be with facial expression, eye gaze [22], non-verbal voice [23], body poseture [24] (head direction [25] , body direction, foot direction), body gesture [26, 27] (sign language if push to the limit), action sequence [28, 29, 30] or even biological signals (accelerated heart or respiratory rate). These non-verbal signals, even not recognized by human selves, convey a high dimension of human state. There is existing work on social perception with observing human action, which is also a non-verbal cue, for a time period to infer human intention and generate assistive plan [14], but they are primarily done in simulation and we designed the robot to catch the nonverbal cues from the human and surrounding environment before a language command is formed and came up with a set of plausible solutions involving visible objects that can be manipulated in the environment.

B. Theory of Mind for Human Robot Interaction

Theory of Mind (ToM) is an innate human ability to infer others' internal states based on external behaviors [31, 32]. Children as young as three years old demonstrate an initial capability in this cognitive function [33, 34]. ToM is typically divided into two categories: Cognitive ToM and Affective ToM. Cognitive ToM pertains to the inference of others' beliefs, intentions, and desires, which are essential for understanding verbal expressions and predicting behaviors in complex social situations [20]. Current HRI frameworks predominantly utilize Cognitive ToM, such as predicting a human's next action or requirement in a rational way [35, 36, 37, 38]. Affective ToM, on the other hand, involves recognizing and responding to the emotional states and feelings of others, facilitating emotional resonance and support [39]. [19] suggested that Affective ToM requires the integration of Cognitive ToM and empathy. Cognitive ToM is crucial for understanding emotional states by inferring the reasons behind someone's emotions or feelings. Empathy is a multifaceted psychological concept encompassing Cognitive, Affective, and Somatic Empathy [40]. Cognitive Empathy refers to the intellectual understanding of another's thoughts and emotions. Affective Empathy involves empathetically resonating with another person's emotional experiences. Somatic Empathy includes the physical sensation of what another individual is experiencing. These aspects of empathy are instrumental in enhancing HRI by enabling robots to better interpret and respond to both the mental and physical states of humans.

III. Problem statement

In addressing the challenge of enabling a robot to accurately identify and satisfy human needs, we adopt a comprehensive process that starts with gathering detailed observations and concludes with the executing specified actions. This process includes visual data collection analysis on humans and the environment, inference of human internal states, identification of needs, task formulation, and the final decomposition of tasks into executable actions.

The process begins with the observation O of human behaviors and environmental factors from the input information I (a RGBD image in our case) using a generative pre-trained model $\mathcal{M}_{\text{observe}}$, respectively denoted as:

$$\begin{aligned} O = \{&O_{\text{human}} = \{o_{\text{f}}, o_{\text{e}}, o_{\text{h}}, o_{\text{g}}, o_{\text{p}}\}, \\ &O_{\text{environment}} = \{o_{\text{l}}, o_{\text{I}}, \{obj_{\text{human}}^{\text{k}}\}_{k=1}^{n_1}\}\} \\ &= \mathcal{M}_{\text{observe}}(I) \end{aligned} \tag{1}$$

where in human observation, o_{f}, o_{e}, o_{h}, o_{g}, and o_{p} denote facial expressions, eye gaze, head direction, gestures, and posture, respectively; in environmental observations, o_{l}, o_{i}, $obj_{\text{human}}^{\text{k}}$ and n_1 denote location, illumination conditions, interactive object k for human, the number of interactive object for human respectively.

From these observations, $\mathcal{M}_{\text{state}}$ infers the internal states (S) of individuals, which includes both mental and physical states such as emotions, intentions, and various sensory inputs:

$$\begin{aligned} S = \{&S_{\text{physical}} = \{s_{\text{t}}, s_{\text{ta}}, s_{\text{v}}, s_{\text{s}}, s_{\text{sm}}, s_{\text{ve}}, s_{\text{p}}, s_{\text{i}}\}, \\ &S_{\text{mental}} = \{s_{\text{em}}, s_{\text{a}}, s_{\text{d}}, s_{\text{in}}\}\} \\ &= \mathcal{M}_{\text{state}}(O) \end{aligned} \tag{2}$$

where in physical states, s_{t}, s_{ta}, s_{v}, s_{s}, s_{sm}, s_{ve}, s_{p}, and s_{i} denote touch, taste, vision, sound, smell, vestibular, proprioception, and interoception respectively; in the mental states, s_{em}, s_{a}, s_{d}, and s_{in} denote emotion, attention, desire, and intention respectively.

These states allow us to identify any unmet needs (N): $N = \mathcal{M}_{\text{needs}}(S)$. Upon identifying these needs, we analyze the need and constraint (C) of the robot body and the environment to generate feasible tasks (T) by$\mathcal{M}_{\text{task}}$, denoted as:

$$T = \mathcal{M}_{\text{task}}\left\{N, C = \{c_{\text{m}}, c_{\text{g}}, c_{\text{o}}, \{obj_{\text{robot}}^{\text{k}}\}_{k=1}^{n_2}, \{obj_{\text{robot}}^{\text{potential, k}}\}_{k=1}^{n_3}\}\right\} \tag{3}$$

where in the robot's body constraints, c_{m}, c_{g} and c_{o} represent mobility, grasp ability and functional operating ability, respectively; and in the the environmental context $obj^{\mathrm{k}}_{\mathrm{robot}}$ represents the visible and manipulable objects for robots, $obj^{\mathrm{potential,\ k}}_{\mathrm{robot}}$ represents the potential existing objects for robots in this specific environment.

Finally, the formulated tasks are decomposed into a sequence of t steps executable robotic actions (A) by $\mathcal{M}_{\mathrm{decompose}}$:

$$A = \mathcal{M}_{\mathrm{decompose}}(T) = \{a_1, a_2, \ldots, a_t\},$$
$$a_t \subseteq \{a_m, a_g, a_o\} \tag{4}$$

where a_m, a_g, a_o denotes moving to a certain target(object or human), grasping a certain object and operating a certain object.

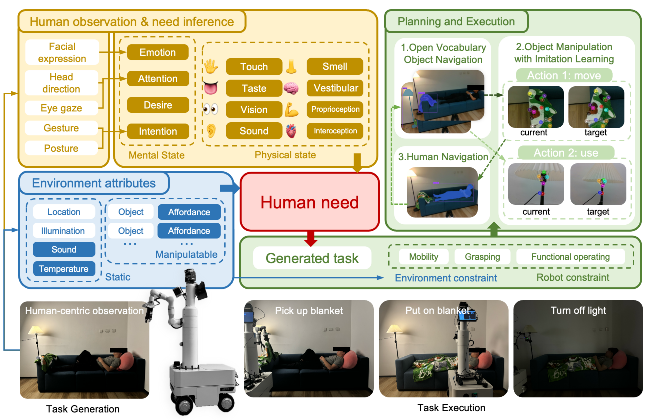

Fig. 2. Overview of AToM-Bot, a robotic system for identifying and responding to human needs. It integrates human observations and environmental attributes to infer human needs. It then generates tasks for a robot by navigating to objects, manipulating them, and assisting human in daily setting. Due to the requirements of blind peer review, the individuals in the images were anonymized.

IV. Method

We present a robot-human interaction framework designed to proactively detect and address human needs empowered by open-vocabulary perception and deduction capabilities of VLM and visual foundation model (See Figure. 3). To be more specific, we employ GPT-4V [41] to (i) extract the observation of human behaviors and environment factors, (ii) infer internal human states from the observations, and (iii) finally generate executable tasks catering to the unexpressed human needs. As for the physical execution of the generated task plan, we rely on Grounding SAM [42, 43, 44] to pinpoint the locations of the concerning objects and human users, and DINO-ViT features [45, 46]to retrieve the knowledge of low-level manipulation skills from pre-recorded human demonstrations.

A. Proactive Human Need Detection

Our robot employs a head-mounted camera to observe humans. Relying solely on raw observations for input into VLM often leads to misinterpretations of human states. *How do human infer another person's internal states based solely on image?* In fact, when we engage in affective social reasoning, it is easy to re-experience and bring past experiences into the process [47, 48]. Recalling these experiences can also activate related sensory pathways, causing similar responses. This capacity for empathetic thinking is precisely what robots lack. Here, we employ reasoning based on common sense, using typical human mental states [49, 50] and physical states [51, 52] as cues to prompt robots to deduce current human feelings through visual clues.

From a first-person viewpoint, the camera captures various non-verbal human expressions such as facial expressions, eye gaze, head direction, gestures, and posture, along with environmental attributes like location, lighting, and nearby objects. These observations enable us to deduce mental states (e.g., emotions, attention, intentions) and physical conditions (e.g., senses of touch, taste, vision, hearing, smell, as well as vestibular, proprioceptive, and interoceptive senses), and also to recognize environmental factors beyond visual cues, such as sound and temperature.

Integrating all these elements, we propose a structured prompt for the VLM that systematically incorporates these diverse data streams to infer human needs, as depicted in Figure. 2. Through a detailed description of relative factors, we bring the subconscious AToM in humans during social inference into our robot, which greatly enhances its ability to understand human internal states.

B. Embodied Task Generation for Human Need Solution

Upon identifying a specific human need (or potentially multiple needs), the robot should devise a practical and immediate solution. Equipped with mobility and the ability to manipulate objects with a single hand, the robot's actions must be feasible and directly executable [1, 12]. Instead of complex tasks like completing someone's work or moving heavy appliances like a fridge, the robot should focus on simpler, achievable tasks within its capabilities and current context.

The robot can interact with objects visible in the scene or infer the presence of everyday items based on the scene, utilizing common sense reasoning. For example, it might predict a yoga block in a gym but not in an office. The robot's manipulative abilities are categorized into two primary actions: moving and using objects. This distinction is crucial due to its single gripper and one-arm configuration.

The VLM evaluates whether fulfilling a need involves merely moving or actively using an object. For instance, transferring a blanket to a person might satisfy a need, whereas merely moving a lamp to the person would be inappropriate and ineffective. If a need involves using an object, the robot must determine whether it can perform the task given its configuration. For example, it can turn on a light but cannot operate a massage gun for a human. This approach ensures that the robot's responses are both relevant and within its operational scope.

C. Open-Vocabulary Mobile Manipulation for Task Execution

Task-Conditioned Navigation Once the robot generates a task to perform, it employs its pan-tilt unit (PTU) to scan the environment dynamically. The scan starts from a central, neutral position and includes systematic horizontal and vertical sweeps, ensuring comprehensive visual coverage. The system uses the Grounding SAM [44] algorithm to segment and identify relevant objects from the visual data. Upon detecting an object that matches the specified criteria, the robot utilizes depth information from RGB-D images to pinpoint the object's location.

Once the position of the object is calculated, the robot plans its navigation, balancing the necessity of proximity for precise manipulation with the importance of maintaining a safe distance. The navigation goal is set at a point along the extended line of the end effector's z-axis, optimizing the approach towards the object.

Additionally, the robot must locate the person, if the identified task involves moving an object to a human. For this, we utilize the Grounding SAM [44] algorithm to detect humans within the environment.

Manipulation via Trajectory Alignment For object manipulation, we follow the DINOBot [46] framework, which utilizes DINO-ViT features to perform the spatial alignment between the current observation and the pre-recorded demonstrations for manipulation skills. When the robot encounters an object, it uses image-level semantic features to identify the most visually similar object from a database of historical human demonstrations. Once a visually similar object is identified, the robot utilizes pixel-level geometric features to precisely align the relative spatial position and orientation between the camera frame and the target object. This alignment, crucial for effective interaction, involves adjusting the robot's position and orientation to align with the pre-recorded trajectory. Once there is an error between the current relative pose and the pre-recorded relative pose, the robot arm executes a delta end-effector path to complete the manipulation task.

V. Experiments

A. Experimental setup

Scenarios: Our experiment setup involved 16 scenarios (Figure. 3) conducted within four commonly encountered everyday environments: office spaces, home gyms, living rooms, and kitchens. In each setting, humans typically encounter various physical or mental states of distress. We displayed 4 typical instances from each environment on the upper half: eating spicy food in the kitchen, feeling tired in the office, sweating while cycling in a home gym, and falling asleep on a sofa. In order to provide more possibilities for the robot to operate and ensure the ecological validity of the test, we also prepared objects matching the scene in the same space outside the picture, ranging from 10 to 18 kinds for each scene. This includes the solution for the robot we tested in advance in Section 5.3.

Human-centered Metrics: We deployed our robot to react to these scenarios and recruited 59 participants to conduct a series of evaluations on satisfaction with various stages of our system and its similarity to the human mind. We also conducted a real-world execution success rate analysis.

Robot Setup: The robot is a mobile manipulator with a 7 degree-of-freedom arm from Realman, a parallel gripper from DH-ROBOTICS, and an omnidirectional mobile base from Agilex. It uses two Realsense 435 cameras, with the top camera running at 640x480, and the wrist camera running at 320x240.

B. Comparison and Evaluation

Human Need Detection In Experiment 1, participants were presented with images depicting various scenarios and asked to provide open-ended responses identifying the inferred needs. Subsequently, they evaluated the robot's responses (output of using prompt described in Section 4.1) on a scale from 1 to 7, where 1 indicates extreme dissatisfaction and 7 indicates extreme satisfaction. We employed text embedding and unsupervised clustering algorithm [53] to categorize the responses into 3-5 distinct needs and corresponding solutions. The need most frequently identified by humans served as the standard for comparing the robot's generated needs, resulting in a Similarity score. The average similarity score was 72.8%. The overall satisfaction rating was 6.42 ± 0.32 out of 7, suggesting that, despite variability in human output, participants were generally delighted with our outputs, and the robot's output is very similar to the average human. The human need detection phase demonstrated a high degree of alignment between the robot's inference of needs and human perceptions.

Embodied Task Generation as Human Need Solution In Experiment 2, participants were asked to generate solutions corresponding to the needs in Experiment 1 and to rate satisfaction with the solutions generated by the robot. The analysis methods were the same as those used in Experiment 1. The average similarity score was 69.6%. The overall satisfaction rating was 6.15 ± 0.55. It is noteworthy that in Tasks 5 and 8, the similarity between the robot's and the human's solutions was relatively low (16.9% and 13.1%), yet the satisfaction ratings remained high(6.52 ± 0.71 and 5.54 ± 1.56). This is because most humans chose to directly use their hands to help stabilize or push the participant for better stretching, an approach unsuitable for our robot. Instead, our robot opted to find assisting equipment within its capability range, such as yoga bricks or resistance bands. This is a prime example of the effectiveness of the embodied constraints in our prompt.

Real World Execution For real-world implementation, we recorded videos of the robot successfully executing solutions from Experiment 2 in a third-person perspective. Participants scored the keyframes for satisfaction, resulting in an average satisfaction rating of 6.17 ± 0.54. We also evaluated the robot's task success rate. Since the tasks were long-horizon, involving various factors such as object detection accuracy, spatial alignment precision with records, and navigation execution accuracy, there were considerable fluctuations in success rates, especially in tasks requiring precise operations like turning on lights or handling soft, whole-body movements like covering with a blanket. Nevertheless, the overall performance was 69.4%. See details in AToM-Bot.

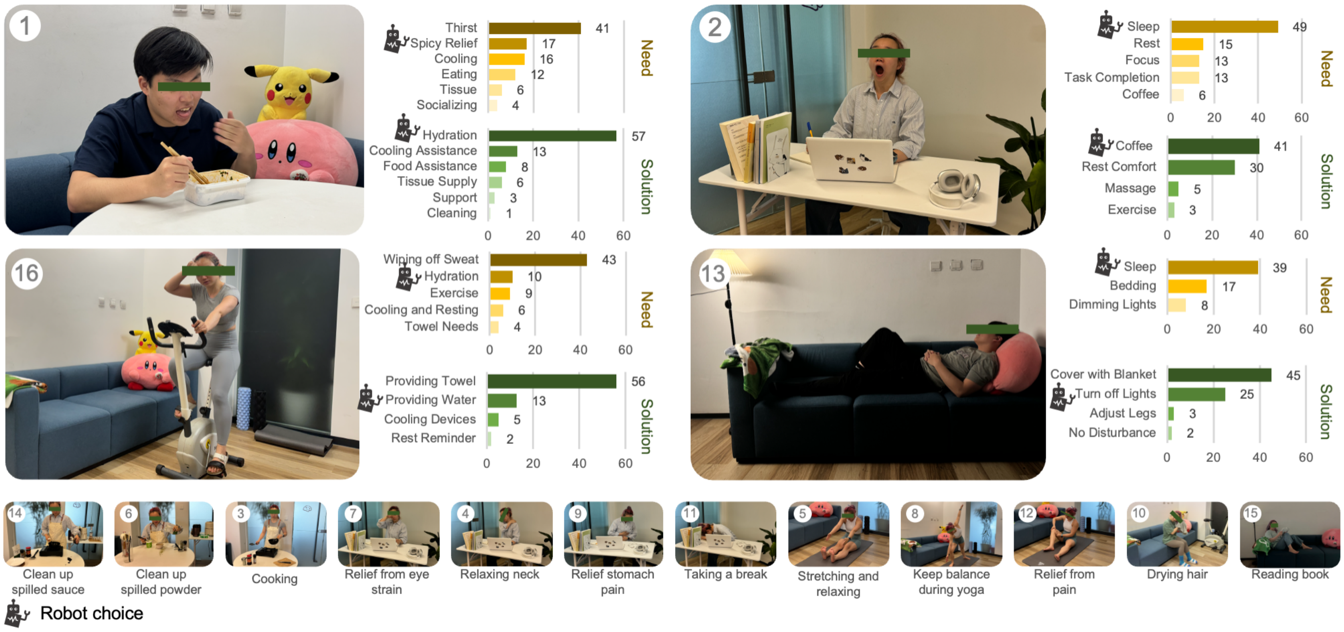

Fig. 3. Human and robot response to 16 task scenarios. The human responses are shown in the bar chart. The robot responses are marked by the "robot choice" icon. Due to the requirements of blind peer review, the individuals in the images were anonymized.

Task ID	Need Detection		Embodied Solution		Task Execution	
	Similarity	Satisfaction	Similarity	Satisfaction	Satisfaction	Success Rate
1	28.8%	6.59 ± 0.62	**96.7%**	6.23 ± 0.98	6.46 ± 0.72	8/10
2	83.1%	6.27 ± 0.86	69.5%	5.42 ± 1.64	5.83 ± 1.33	9/10
3	88.1%	5.83 ± 1.15	39.0%	5.83 ± 1.42	6.25 ± 0.93	5/10
4	78.0%	6.58 ± 0.62	78.0%	6.63 ± 0.51	6.24 ± 1.07	8/10
5	49.2%	6.55 ± 0.69	16.9%	6.52 ± 0.71	6.52 ± 0.83	4/10
6	76.3%	6.63 ± 0.68	98.3%	5.39 ± 1.97	5.52 ± 1.66	10/10
7	**94.9%**	6.61 ± 0.57	83.1%	6.59 ± 0.69	6.42 ± 0.91	4/10
8	88.1%	5.79 ± 1.47	13.1%	5.54 ± 1.56	6.31 ± 1.14	10/10
9	30.5%	6.45 ± 0.77	35.5%	6.52 ± 0.74	6.11 ± 1.28	5/10
10	84.7%	6.68 ± 0.55	89.1%	$\mathbf{6.76 \pm 0.49}$	6.79 ± 0.48	10/10
11	83.1%	6.55 ± 0.77	48.0%	5.37 ± 1.91	4.96 ± 1.82	5/10
12	89.8%	6.62 ± 0.59	88.1%	6.65 ± 0.70	6.42 ± 0.94	9/10
13	69.5%	6.48 ± 0.86	76.2%	6.37 ± 1.12	$\mathbf{6.82 \pm 0.52}$	4/10
14	64.4%	6.66 ± 0.58	93.2%	6.55 ± 0.83	6.34 ± 0.92	7/10
15	83.1%	$\mathbf{6.70 \pm 0.61}$	94.9%	6.69 ± 0.68	6.55 ± 0.75	4/10
16	72.9%	5.82 ± 1.40	94.9%	5.78 ± 1.68	5.91 ± 1.55	9/10
	72.8%	6.42 / 7	69.6%	6.15 / 7	6.17 / 7	69.4%

TABLE I
HUMAN COMPARISON AND EVALUATION OF NEED DETECTION, EMBODIED SOLUTION, AND TASK EXECUTION OF ATOM-BOT.

VI. CONCLUSION

In this work, we introduce **AToM-Bot**, a comprehensive framework for inferring human mental and physical needs, utilizing Vision Language Models enhanced with Affective Theory of Mind. AToM-Bot proactively generates and executes tasks based on these inferences, tailored to real-world environments and everyday scenarios. This approach provides substantial benefits in adapting to a wide range of unspoken instructions and dynamic human contexts, enabling more intuitive and effective robot-human interactions.

REFERENCES

[1] 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. 2023. Publisher: arXiv Version Number: 2.

[2] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models, September 2022. arXiv:2209.11302 [cs].

[3] Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *ArXiv*, January 2022.

[4] 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, September 2023. arXiv:2307.01928 [cs, stat].

[5] Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. Translating Natural Language to Planning Goals with Large-Language Models. 2023. Publisher: arXiv Version Number: 1.

[6] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah

Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models. arXiv, 2022. Version Number: 1.

[7] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An Embodied Multimodal Language Model. arXiv, 2023. Version Number: 1.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 2023. Publisher: arXiv Version Number: 1.

[9] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. ChatGPT Empowered Long-Step Robot Control in Various Environments: A Case Application. *IEEE Access*, 11:95060–95078, 2023.

[10] Marta Skreta, Naruki Yoshikawa, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Errors are Useful Prompts: Instruction Guided Task Programming with Verifier-Assisted Iterative Prompting. 2023. Publisher: arXiv Version Number: 1.

[11] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2Motion: From Natural Language Instructions to Feasible Plans. *Autonomous Robots*, 47(8):1345–1365, December 2023. arXiv:2303.12153 [cs].

[12] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, K. Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, A. Irpan, Eric Jang, Rosario M. Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, N. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, S. Levine, Yao Lu, Linda Luu, Carolina Parada, P. Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, D. Reyes, P. Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. April 2022.

[13] 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. arXiv, 2023. Version Number: 2.

[14] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration, May 2021. arXiv:2010.09890 [cs].

[15] Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart Help: Strategic Opponent Modeling for Proactive and Adaptive Robot Assistance in Households, April 2024. arXiv:2404.09001 [cs].

[16] Yuening Zhang. *Effective Teamwork Using a Theory of Mind Over Plans*. Thesis, Massachusetts Institute of Technology, February 2024. Accepted: 2024-03-21T19:13:43Z.

[17] Lance Ying, Kunal Jha, Shivam Aarya, Joshua B. Tenenbaum, Antonio Torralba, and Tianmin Shu. GOMA: Proactive Embodied Cooperative Communication via Goal-Oriented Mental Alignment, March 2024. arXiv:2403.11075 [cs].

[18] OpenAI. gpt4v. https://openai.com/contributions/gpt-4v/, 2023.

[19] Simone G Shamay-Tsoory, Hagai Harari, Judith Aharon-Peretz, and Yechiel Levkovitz. The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, 46(5):668–677, 2010.

[20] Simone G Shamay-Tsoory and Judith Aharon-Peretz. Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia*, 45(13):3054–3067, 2007.

[21] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv: 2403.12910*, 2024.

[22] Suleyman Ozdel, Yao Rong, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. Gaze-Guided Graph Neural Network for Action Anticipation Conditioned on Intention, April 2024.

[23] Kristina T. Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W. Picard. ReCANVo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523, August 2023.

[24] Yu Luo, Jianbo Ye, Jr Adams, Jia Li, Michelle G. Newman, and James Z. Wang. ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild. *International Journal of Computer Vision*, 128(1):1–25, January 2020.

[25] Jialong Xie, Botao Zhang, Qiang Lu, and Oleg Borisov. A Dynamic Head Gesture Recognition Method for Real-time Intention Inference and Its Application to Visual Human-robot Interaction. *International Journal of Control, Automation and Systems*, 22(1):252–264, January 2024.

[26] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots, October 2018.

[27] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. SMG: A Micro-gesture Dataset Towards Spontaneous Body Gestures for Emotional Stress State Analysis. *International Journal of Computer Vision*, 131(6):1346–1366, June 2023.

[28] Lei Shi, Paul-Christian Bürkner, and Andreas Bulling. Inferring Human Intentions from Predicted Action Probabilities, March 2024.

[29] Maithili Patel, Aswin Prakash, and Sonia Chernova. Predicting Routine Object Usage for Proactive Robot Assistance, January 2024.

[30] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home Action Genome: Cooperative Compositional Action Understanding.

[31] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[32] Ian A Apperly and Stephen A Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953, 2009.

[33] John H Flavell, Frances L Green, and Eleanor R Flavell. Developmental changes in young children's knowledge about the mind. *Cognitive Development*, 5(1):1–27, 1990.

[34] Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. 1992.

[35] Huaxiaoyue Wang, Kushal Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, Gonzalo Gonzalez-Pumariega, Aditya Kompella, Maximus Adrian Pace, Yash Sharma, Xiangwan Sun, Neha Sunkara, and Sanjiban Choudhury. Mosaic: A modular system for assistive and interactive cooking, 2024.

[36] Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart Help: Strategic Opponent Modeling for Proactive and Adaptive Robot Assistance in Households, April 2024.

[37] Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling, 2022.

[38] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration, 2021.

[39] Dandan Li, Xiaosi Li, Fengqiong Yu, Xingui Chen, Long Zhang, Dan Li, Qiang Wei, Qing Zhang, Chunyan Zhu, and Kai Wang. Comparing the ability of cognitive and affective theory of mind in adolescent onset schizophrenia. *Neuropsychiatric Disease and Treatment*, pages 937–945, 2017.

[40] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. Empathy: A review of the concept. *Emotion review*, 8(2):144–153, 2016.

[41] OpenAI. gpt4v. https://openai.com/contributions/gpt-4v/, 2023.

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[44] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[45] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021.

[46] Norman Di Palo and Edward Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[47] Josef Perner, Daniela Kloo, and Edith Gornik. Episodic memory development: Theory of mind is part of re-experiencing experienced events. *Infant and Child Development: An International Journal of Research and Practice*, 16(5):471–490, 2007.

[48] Vittorio Gallese. Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):659–669, 2007.

[49] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.

[50] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.

[51] E Bruce Goldstein and James R Brockmole. *Sensation and perception*, volume 90. Wadsworth-Thomson Learning Pacific Grove, CA, USA., 2002.

[52] J Brendan Ritchie and Peter Carruthers. The bodily senses. *The Oxford handbook of philosophy of perception*, pages 353–370, 2015.

[53] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classifica-

tion. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[54] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation, 2022.

[55] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation, 2023.

VII. Appendix

A. Ablations

We performed ablation studies to assess the impact of key components: AToM for need detection and the embodied constraints for grounding task generation in our prompt. These components were evaluated for their contribution to need similarity (Table II) and solution similarity (Table III) , detailed in Appendix. **AToM:** Removing AToM significantly lowers need and solution similarity by 26.4% and 31.2%, respectively. **Embodied Constraints:** Operating without constraints results in much lower scores in both need and solution similarity by 4.1% and 38.6%. Employing both AToM and embodied constraints as AToM-bot achieves the highest improvements in need and solution similarity, reaching 72.8% and 69.6%. These results confirm that both AToM and embodied constraints are essential for aligning with human reaction. The ablation studies further solidified the importance of both AToM for accurate need detection and the role of embodied constraints in ensuring the practicality of generated tasks.

These results highlight AToM-Bot's capability to not only understand and predict human needs effectively but also to generate and execute tasks that address these needs in a real-world setting.

Prompt	w/o AToM	w/ AToM
w/o Constraints	33.2%	68.7%
w/ Constraints	46.4%	**72.8%**

TABLE II
Need similarity.

Prompt	w/o AToM	w/ AToM
w/o Constraints	4.9%	31.0%
w/ Constraints	38.4%	**69.6%**

TABLE III
Solution similarity.

B. Limitation and Future Work

Despite compelling results, AToM-Bot has several limitations. First, the robot is equipped with only two cameras, limiting its sensory input to visual information. In future enhancements, the robot could be equipped with multi-modal sensors like audio sensors and thermometers, enabling it to receive more information and analyze human needs with greater breadth. Second, the current experiment did not employ sequential detection; long-term dynamic observations rather a snapshot will further increase the detection precision. Third, all the target objects are within the robot's visibility by turning the PTU. If equipped with zero-shot object navigation modules [54, 55] the robot will have the potential to navigate outside the room for more complex task and more possible solutions. Fourth, we observed cases where human participants' solutions differed from those of the robot in solutions to eating spicy food and stomach discomfort that predominantly resulted from cultural backgrounds. Future involvement of participants from diverse cultural backgrounds will provide greater diversity. Lastly, although our framework exhibits a certain degree of generalizability, where similar visual features of objects allow for the transfer of recorded trajectories—such as among different cups or between handheld vacuum and hair dryers—these capabilities are still constrained by our reliance on pre-recorded movements. In the future, we could enhance the robot's flexibility by exploring more trained end-to-end models that extend from visual features to low-level actions, thereby developing more adaptable and autonomous movement strategies.