C2ST
C2 Science and Technology

Research Article
https://doi.org/10.26599/C2ST.2026.9700006

Situated proactive task generation for embodied agents

Haifeng Xu[1], Chunwen Li[1], Xiaohu Yuan[2], Huaping Liu[3,*]

[1] Department of Automation, Tsinghua University, Beijing 100084, China
[2] Qiyuan Laboratory, Beijing 100095, China
[3] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ABSTRACT

Embodied agents operating in everyday environments must go beyond explicit instruction following toward proactive assistance, in which they infer latent human needs from ongoing dialogue, ground these needs in the physical scene, and determine whether and how to act. However, existing vision-language-action systems are primarily designed for explicit task execution and struggle to jointly reason over conversational context and dynamic embodied environments. In this work, we introduce HomeSPT-Bench, a benchmark for situated proactive task generation built on the OmniGibson simulator, in which agents must infer implicit needs from multi-party dialogue, ground them in fully specified 3D scenes, and decide on appropriate actions. Evaluations show that state-of-the-art VLM-based agents exhibit a substantial performance gap in this setting, with frequent failures in spatial grounding, occlusion handling, and intervention decision-making. HomeSPT-Bench provides a testbed for studying situated proactive behavior that integrates dialogue understanding with embodied scene grounding, and highlights the remaining gap between language-based reasoning and reliable embodied proactive assistance.

KEYWORDS

proactive robot; embodied AI; situated task generation; dialogue-driven task inference; interactive embodied agent

1 Introduction

Robots are increasingly being deployed from structured laboratory settings into everyday human environments such as homes, offices, hospitals, and hotels. In these service contexts, users expect embodied agents to move beyond executing pre-defined commands and instead to understand ongoing situations, recognize emerging needs, and provide timely assistance. However, most existing robot systems still follow an explicit-command paradigm: a user specifies a goal, the robot executes a corresponding skill, and the interaction terminates upon completion. While effective for benchmarking and controlled tasks, this paradigm does not reflect how human needs naturally arise in daily life, where requests are often implicit, indirect, or embedded within conversation and context.

This limitation motivates the development of proactive embodied agents, which continuously reason about their environment and human activities rather than passively waiting for explicit instructions. Prior work has explored various aspects of proactivity, including when to intervene, how to infer human intent, and how to anticipate routine activities[1-3]. Early non-linguistic approaches infer intent from physical cues such as gaze direction, body orientation, object trajectories, or human-object interactions[4-7]. These studies show that proactive assistance requires more than detecting a possible intent: an agent must also decide whether intervention is appropriate, feasible, and beneficial in the current situation. This requirement is especially important because unnecessary or poorly timed assistance may disrupt users, reduce their autonomy, or trigger inappropriate actions.

Natural language dialogue provides a rich source of evidence for inferring latent human needs. In everyday environments, humans rarely issue direct commands to robots; instead, they often express needs implicitly through conversation, such as by commenting on environmental conditions, missing objects, safety concerns, or upcoming activities. Recent studies have begun to exploit this signal for proactive task inference. For example, autonomous navigation task generation from natural language dialogue studies how a robot can infer office navigation goals from multi-person conversations[8], while ProactiveDialog introduces simulated proactive dialogue data for robot task inference from streaming workplace conversations[9]. These works demonstrate that multi-party dialogue can reveal task intent even without explicit robot-directed commands. However, the inferred tasks are mainly associated with office delivery or workplace scenarios, and the physical scene does not substantially reshape the meaning of the task across different embodied states.

In parallel, embodied AI has made substantial progress in building agents that perceive, reason, and act within simulated environments. From the perspective of embodied intelligence, intelligent behavior emerges from the tight coupling among morphology, action, perception, learning, and the surrounding environment[10]. This view motivates embodied benchmarks that evaluate agents not only as language reasoners, but also as situated systems that interact with and adapt to physical contexts. Interactive platforms such as AI2-THOR[11], Habitat[12–13], VirtualHome[14], ProcTHOR[15], and BEHAVIOR-1K with OmniGibson[16] provide

Received: June 21, 2026; Revised: July 12, 2026; Accepted: August 5, 2026

*Huaping Liu, hpliu@tsinghua.edu.cn

Citation: Xu H F, Li C W, Yuan X H, et al. Situated proactive task generation for embodied agents[J]. *C2 Science and Technology*, 2026, 1(1): 83-90. https://doi.org/10.26599/C2ST.2026.9700006.

© Chinese Institute of Command and Control. Publishing services by Tsinghua University Press. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

increasingly diverse, physically grounded, and reproducible environments for studying navigation, manipulation, activity understanding, and long-horizon household tasks. On top of these environments, embodied benchmarks have progressed from point-goal navigation to language-conditioned instruction following and interactive reasoning. ALFRED[17] requires agents to complete household chores from free-form English instructions; vision-and-language navigation benchmarks such as Room-to-Room[18–19] evaluate language grounding during indoor navigation; Embodied question answering requires agents to explore a scene before answering a question[20–21]; and TEACh introduces human-human dialogues in which an agent must extract task-relevant intent from unconstrained multi-turn conversation before acting[22].

Recent advances in large language models (LLMs), vision-language models (VLMs), and vision-language-action models (VLAs) have further expanded the capabilities of embodied agents[23-26]. These models integrate visual perception, language understanding, and action generation, enabling robots to follow natural language instructions across diverse tasks and embodiments. Nevertheless, most existing embodied benchmarks still assume that the agent is given an explicit goal, whether in the form of an instruction, a question, or a dialogue in which the task objective is eventually specified. As a result, the central challenge is usually grounding and execution after task specification, rather than the earlier and more ambiguous stage of determining whether a task exists, what the task should be, and whether the agent should intervene at all.

More recent proactive assistance studies have begun to address this gap, but important limitations remain. Smart Help[27] models proactive and adaptive household support by reasoning about both the human's goal and the human's ability to complete it independently. Attentive Support[28] studies how a robot can support human groups without disturbing them by combining scene perception, dialogue acquisition, situation understanding, and LLM-based reasoning. ProVox[29] explores situated human-robot collaboration, where an LLM-based planner infers user intent from early interactions and proactively suggests helpful actions before explicit instructions. At the level of LLM-based agents, Proactive Agent[30] investigates how agents can anticipate and initiate tasks from ambient user context. RoboOmni[31] further introduces multimodal proactive manipulation by using speech, environmental audio, and visual observations to infer latent user intent. These works collectively highlight a growing consensus that proactive agents should not merely act once a possible need is detected, but should judge whether assistance is needed, appropriate, and grounded in the surrounding context.

However, an important gap remains in how proactive task generation is formulated. Dialogue-centered approaches emphasize conversational cues and implicit intent, but often under-specify the role of the physical environment. Embodied benchmarks emphasize perception, navigation, and manipulation, but usually assume that task goals are already provided. Multimodal proactive systems introduce richer contextual signals, but are often evaluated in localized manipulation scenarios rather than complex environments requiring reasoning over multiple rooms, object states, hidden or occluded objects, and no-action decisions. Consequently, existing research has not yet fully captured the situated nature of proactive assistance: the same dialogue may require different responses depending on the embodied scene.

In this paper, we claim that proactive assistance is fundamentally situated. A conversational cue alone is insufficient to determine the correct action. For example, an expression of discomfort in a room may or may not require intervention depending on whether the environmental condition has already been addressed. Similarly, an object-related need depends on whether the relevant object exists in the scene, where it is located, whether it is accessible, and whether acting on it would be appropriate. Therefore, proactive decision-making must jointly reason over dialogue and embodied context, rather than relying on language alone.

To study this problem, we introduce HomeSPT-Bench, a benchmark for situated proactive task generation in embodied environments. Given a multi-party dialogue and a fully specified simulated scene, an agent must infer whether an implicit need exists, ground the need in the current environment, and determine an appropriate outcome, including the option of taking no action. Unlike conventional instruction-following benchmarks, HomeSPT-Bench does not provide explicit task definitions; instead, the task structure emerges from the interaction between dialogue and environment.

HomeSPT-Bench is constructed using BEHAVIOR-1K household scenes within the OmniGibson simulator[16]. Each instance includes a multi-party dialogue, scene context, object states, and defined success conditions. The benchmark spans six categories of proactive assistance: object fetching, state checking, environmental adjustment, rearrangement, safety-related intervention, and no-action decisions. By grounding task-critical variables such as objects, states, and outcomes in the simulator before dialogue synthesis, HomeSPT-Bench provides consistent and reproducible scenarios for evaluating situated proactive reasoning.

We evaluate state-of-the-art VLM-based agents under three settings with increasing environmental access: dialogue-only reasoning, dialogue with structured scene text, and an active semantic agent operating in a closed loop with the environment. The results show that current models struggle to reliably infer implicit needs and ground them in physical contexts. In particular, they frequently fail in state verification, spatial grounding, occlusion handling, and deciding when no intervention is needed. These findings reveal a substantial gap between language-based reasoning and reliable embodied proactive assistance.

The main contributions of this work are as follows:

1) A new problem formulation, situated proactive task generation, in which the proactive task is determined jointly by multi-party dialogue and the embodied scene state, explicitly targeting the pre-execution stage of task understanding.

2) A simulator-grounded benchmark, HomeSPT-Bench, featuring multi-party dialogues, object states, and six categories of proactive assistance tasks across diverse household scenes.

3) Evaluations across model backbones and interaction settings, demonstrating that situated proactive assistance remains a significant challenge for current LLMs/VLMs, with identified failure modes including spatial grounding errors, occlusion-induced perceptual failures.

2 HomeSPT-Bench

This section introduces HomeSPT-Bench, a dataset and benchmark for situated proactive task generation in embodied environments. The goal of the benchmark is to evaluate whether an embodied agent can infer latent human needs from multi-party dialogue and ground them in the observable embodied environment. Unlike instruction-following datasets, HomeSPT-Bench does not provide the agent with an explicit task command. Instead, each instance specifies a situated scenario in which the task must be inferred from the relation between dialogue content and the embodied environment. Fig. 1 formalizes the research problem in this work: an agent observes an ongoing conversation, grounds the implied need in the surrounding scene, and then

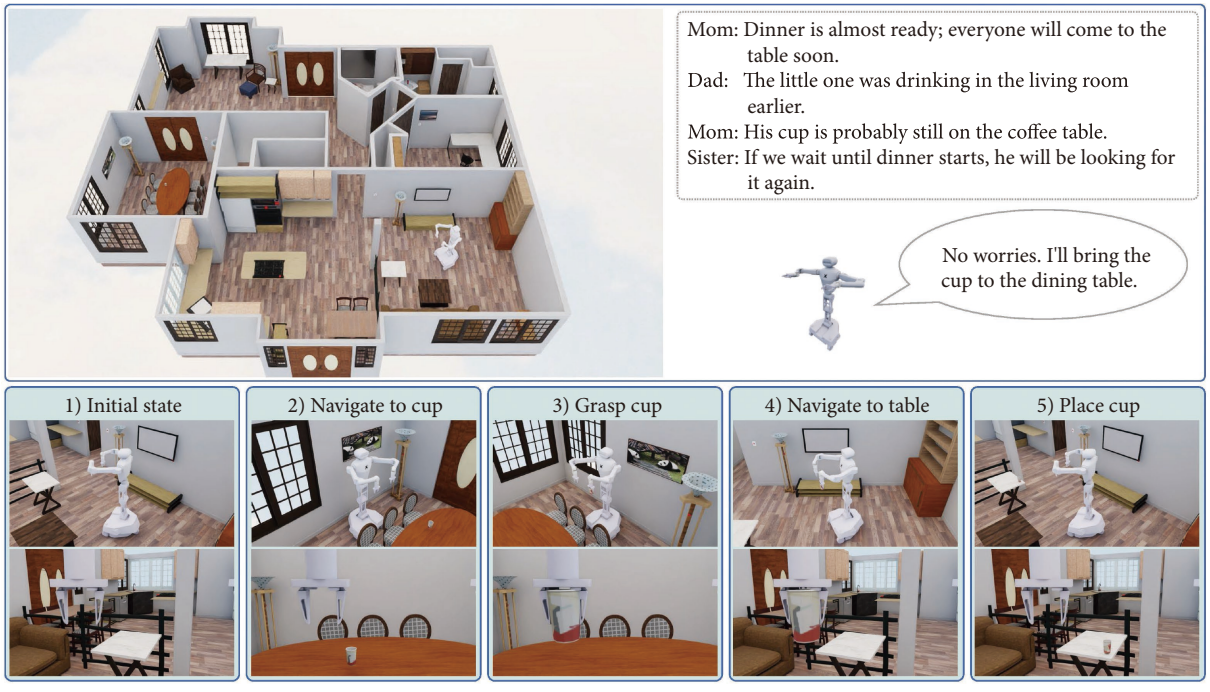


Figure 1 Overview of HomeSPT: situated proactive task generation, where an embodied agent observes family dialogue, identifies a latent need, and grounds the task proactively in the surrounding home environment before acting.

determines whether and how to intervene.

2.1 Problem setup

We define situated proactive task generation as the problem of inferring a grounded proactive task from an indirect dialogue and a situated embodied environment. Formally, each instance is represented as

$$x = (\mathcal{S}, D, r_D, C, Y) \tag{1}$$

where $\mathcal{S}$ is a configured embodied scene instance, D is a multi-party dialogue, r_D is the room where the dialogue occurs, C is the scene context exposed to the model, and Y is the target outcome of the proactive task. Here, $\mathcal{S}$ includes the scene layout, objects, object locations, and initial object states. Different scene configurations may therefore pair with similar dialogues but require different target outcomes.

The target outcome Y describes the desired post-task situation rather than a particular action sequence. For example, Y may specify that a window should be open, an appliance state should be reported, a target object should be located on a support surface, or no environmental change should be made. The model is expected to infer a grounded target

$$\hat{Y} = f(D, r_D, C) \tag{2}$$

which can then be judged against Y directly or executed by a downstream agent policy under a chosen evaluation protocol. Importantly, this formulation does not assume a particular control interface. A system may use high-level semantic actions, low-level motor control, a planner-controller hierarchy, or any other execution mechanism. HomeSPT-Bench defines the situated proactive task and its desired outcome; the action representation belongs to the evaluated method.

The central challenge is that Y is not determined by dialogue alone. The same dialogue may correspond to different target outcomes in different scenes or under different object states. Therefore, a model must answer three coupled questions: whether the dialogue implies a latent need, which concrete scene entities are relevant, and what final state or response would satisfy the need.

2.2 Environment and scenes

HomeSPT-Bench is built on BEHAVIOR-1K household scenes within the OmniGibson simulator[16]. These scenes provide realistic indoor layouts, room types, object categories, support surfaces, articulated objects, and object states. We select 12 representative embodied environments to cover diverse home layouts and everyday activity spaces, as shown in Fig. 2.

The selected scenes contain common household room types such as kitchens, living rooms, bedrooms, child rooms, bathrooms, dining rooms, playrooms, and television rooms. These room types are important because dialogue interpretation is often location-dependent. For example, a conversation about discomfort in a living room may imply ventilation or lighting actions, while a similar conversation in a kitchen may involve appliances, food, or safety-related states. Full household layouts are retained so that the agent may face distractor rooms, repeated object categories, and alternative plausible targets.

2.3 Task definition

HomeSPT-Bench defines a proactive task as a latent need that emerges during dialogue and becomes fully specified only after grounding in the scene context. Each task consists of four levels of information. First, the dialogue provides a natural-language cue, such as a complaint, request, reminder, concern, or casual statement. Second, the dialogue location specifies where the interaction takes place and helps disambiguate the relevant room. Third, the scene describes the available objects, their states, and affordances, which determine feasible actions and outcomes. Fourth, the target outcome specifies the desired post-task state or final response. Figures 3 and 4 illustrate two cases in which identical dialogue leads to different grounded outcomes due to variations in scene state or object availability.

The dataset-level task definition is independent of any specific agent action space. For example, the target outcome may require



Figure 2 Representative embodied environments used in HomeSPT-Bench, which covers diverse layouts and common domestic activity spaces.

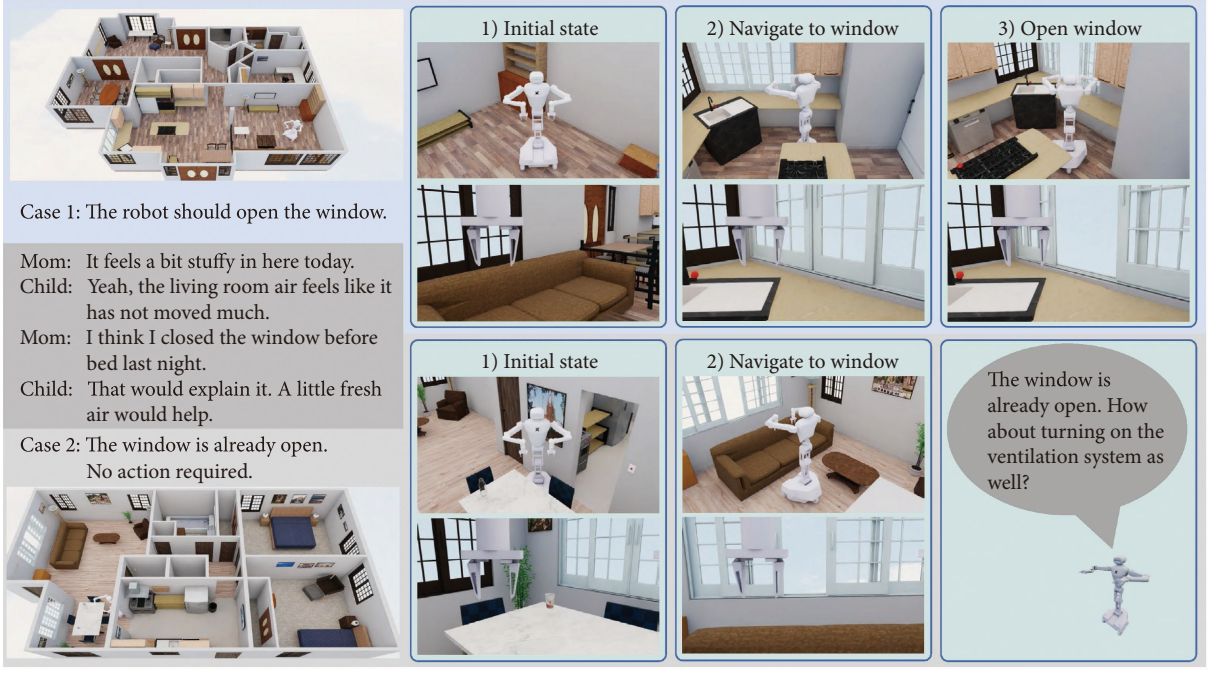


Figure 3 A state-dependent example, where the same dialogue yields different required actions conditioned on the scene state.

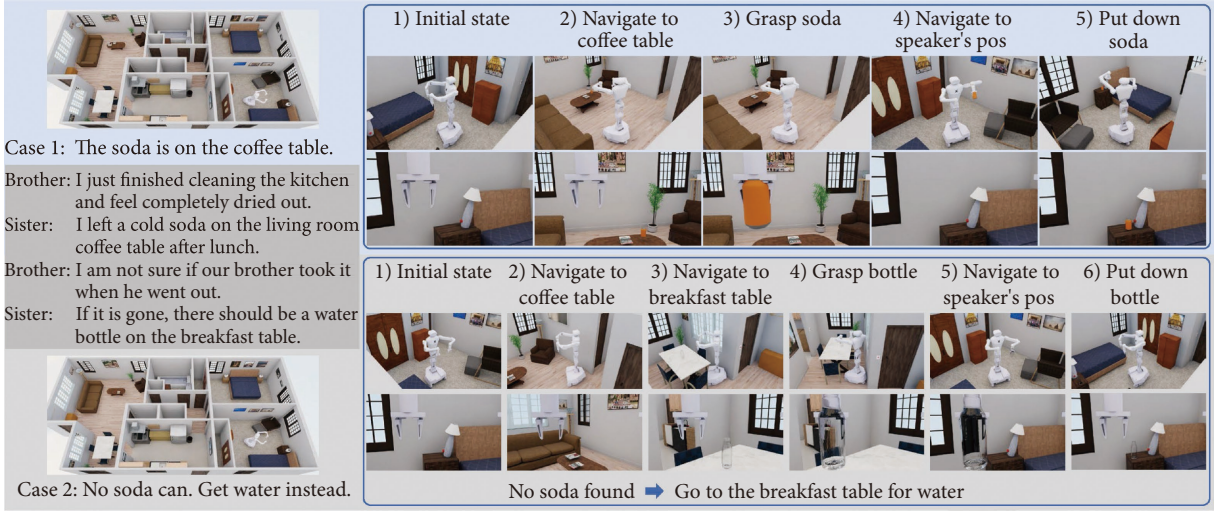


Figure 4 An object-grounding example under identical dialogue, where the same dialogue leads to different executable tasks depending on the scene state and object availability.

that a window be open upon completion, that the agent correctly reports the state of an oven, that a bottle is placed on a designated support, or that the scene remains unchanged when no intervention is necessary. How the agent achieves this outcome is left to

the evaluated method. This separation enables HomeSPT-Bench to support diverse agent embodiments and control abstractions, including high-level planners, semantic action agents, and future low-level control systems.

2.4 Task categories

We design six task categories that cover common proactive assistance scenarios in embodied environments. Each category evaluates a different combination of dialogue inference, scene grounding, state verification, and action selection.

The six categories include object fetching, state checking, environment adjustment, object rearrangement, safety assistance, and no-action cases. Object fetching tasks require the agent to identify a needed item and deliver it to an appropriate target location. State checking tasks require the agent to inspect an object or the state of a room and report the result, without necessarily modifying the environment. Environment adjustment tasks involve modifying environmental conditions related to comfort or functionality, such as ventilation, lighting, or appliance states. Object rearrangement tasks require relocating an object from an unsuitable position to a more appropriate support or container. Safety assistance tasks involve identifying and addressing risky objects or hazardous states, such as unsafe object placement or appliances that should be turned off. No-action tasks evaluate whether the agent can avoid unnecessary intervention when the inferred need is already satisfied, inappropriate, or infeasible.

2.5 Dataset construction pipeline

HomeSPT-Bench is constructed through a structured, multi-stage generation pipeline that decouples simulator-critical task specification from natural language realization, following the general paradigm of structured dialogue and task generation pipelines introduced in ProactiveDialog[9]. The pipeline consists of three sequential stages: scene sampling, structured task generation, and dialogue synthesis, followed by automatic validation and human verification.

Scene and object sampling. We first construct a grounded scene context by sampling a scene instance from an offline inventory derived from BEHAVIOR-1K and OmniGibson. Each scene defines the room layout, object categories, object instances, and candidate support surfaces. For each instance, we further sample a target room and a constrained subset of relevant objects, forming a compact object pool that reflects the local environment while preserving the global scene context.

Structured task generation. Conditioned on the sampled scene and object pool, we prompt GPT-4o to instantiate a task specification. Depending on the task category, the generator selects task-relevant objects, object states, affordances, and, when applicable, source and target supports. It then defines the underlying task objective, including desired object configurations or state changes, expressed in a structured representation consistent with the simulator. Importantly, this stage produces all simulator-critical variables, including object identities, initial states, and target outcomes, which remain fixed in subsequent stages.

Dialogue synthesis. Given the structured task specification and the associated scene context, we prompt GPT-4o to simulate a natural multi-party dialogue between multiple participants. The dialogue is grounded in the sampled environment and reflects the contextual cues, object states, and latent needs defined in the task specification. We generate conversations with 3-8 turns to balance linguistic diversity and task relevance.

This process produces a realistic interaction in which participants gradually express, infer, or respond to potential needs in an implicit manner, forming the basis for downstream task grounding.

Validation and filtering. All generated instances undergo automatic consistency checks to ensure alignment between scene state, task specification, and language description. Invalid samples are discarded and regenerated. In addition, each instance is manually reviewed by human annotators to verify physical plausibility, task solvability, and grounding correctness. Only validated instances are included in the final dataset.

We construct HomeSPT-Bench across 12 household scenes and 6 task categories, with 20 instances per scene–category combination, resulting in a total of 1,440 task instances. All GPT-4o-generated outputs are subject to iterative filtering and regeneration, ensuring that only fully verified and consistent instances are retained.

3 Experimental

This section evaluates HomeSPT-Bench under varying levels of available information. We analyze how current VLM/LLM backbones perform when proactive tasks require dialogue understanding, scene grounding, and active interaction. We compare three evaluation settings with progressively richer environmental access and evaluate each setting using two model backbones: Qwen3.7-Plus and GPT-5.

3.1 Evaluation settings

We consider three evaluation settings: Dialogue-only, Dialogue + scene text, and Active semantic agent. These settings share the same output interface and are evaluated across the same task categories, while differing in the amount and modality of information available to the model.

Dialogue-only. The dialogue-only setting provides the model with only the multi-party dialogue and interaction history. No structured scene information or visual observations are provided. This setting tests whether the model can infer latent user needs from language alone. However, without access to the environment, the model may generate plausible intentions while failing to determine object availability, verify whether target states are already satisfied, or assess whether intervention is required.

Dialogue + scene text. The dialogue + scene text setting augments the dialogue with structured textual representations of the scene. This includes the dialogue location, room types, object categories, available entities under the current setting, and task-relevant state information. This setting evaluates whether explicit symbolic environmental context improves grounding for proactive task inference. Compared with dialogue-only setting, it provides additional evidence about the environmental situation, but remains a static reasoning setup without closed-loop interaction.

Active semantic agent. The active semantic agent setting evaluates the model as an interactive embodied agent operating in a step-by-step decision loop. At each step, the model receives the dialogue, scene text, action history, last action feedback, and optionally a first-person observation, and outputs a high-level semantic action. The environment then executes the action and returns updated feedback for the next step. This setting evaluates whether the model can actively gather information, update its understanding of the situation, select appropriate actions, and terminate when the task is completed. The action space follows the high-level semantic interface defined in Table 1.

3.2 Metric

We report episode success rate as the primary evaluation metric.

Table 1 High-level semantic action interface for the active evaluation setting, which abstracts low-level execution and emphasizes semantic decision making

Action	Description
navigate(x)	Move to a room or object
inspect(x)	Observe or query an object, support, or room
open(x)	Open an object such as a window or fridge
close(x)	Close an object such as a fridge or cabinet
turn_on(x)	Turn on a switch, lamp, or appliance
turn_off(x)	Turn off a switch, lamp, or appliance
pick(x)	Pick up a target object
place(x, y)	Place object x on or inside target y
answer(t_{text})	Provide a final response
no_op()	Intentionally take no action

An episode is considered successful if the final behavior satisfies the target outcome specified in the benchmark instance. Depending on the task family, success requires achieving a target object relation, reaching a desired object state, correctly reporting a state, or correctly determining that no intervention is necessary. We further report success rates across task families to analyze how different information settings affect distinct types of proactive household tasks.

4 Results and discussion

4.1 Results

Table 2 presents the results across six task families. The results show that the two models perform comparably on our benchmark. Both models exhibit weak performance under the dialogue-only and scene text settings, whereas the active semantic setting enables partial success on certain tasks, allowing the models to achieve task objectives to some extent.

We further observe substantial variation in task difficulty across categories. In particular, the rearrangement and safety assistance tasks are significantly more challenging, as they require stronger capabilities in both perception and action execution.

The poor performance in the no-action setting can be attributed to the lack of explicit constraints on proactivity. As a result, the agent may become overly proactive and generate unnecessary actions in order to avoid missing potential task opportunities.

Figures 3 and 4 illustrate representative examples of task generation and execution in situated scenes. These examples demonstrate that identical dialogues can lead to different executable behaviors depending on the current environment state. Figure 3 shows a state-dependent case. Given the same dialogue script, the required action varies with the household state. When the window is closed, the agent grounds the instruction into a concrete task and executes the required action. When the window is already open, no physical intervention is required, and the agent correctly infers that no action should be taken. Figure 4 presents an object-grounding case under identical dialogue. The same inferred household intent is mapped to different executable tasks depending on scene state and object availability. When a coke is available on the coffee table, the agent performs a direct retrieval-and-delivery of the coke to the user. When the coke is not present in the scene, the agent instead adapts to available objects and retrieves a water bottle from the breakfast table for delivery.

4.2 Discussion

We also observe several representative failure cases under the active semantic setting, as illustrated in Figs. 5 and 6. These examples highlight two common limitations of current VLM-based agents in embodied environments.

Figure 5 shows a failure case in cross-room grounding, where the agent incorrectly associates the target object with an irrelevant room and consequently searches an incorrect bookcase. This behavior indicates that the model struggles with consistent spatial grounding across distinct rooms, revealing weaknesses in holistic scene understanding and long-horizon grounding. Figure 6 presents a failure case caused by insufficient visual evidence under occlusion. The target object is overlooked due to partial visibility, suggesting that the agent relies too heavily on immediate perceptual cues and lacks robust active perception strategies for disambiguating partially observed objects.

Overall, evaluations of both failure modes suggest that directly applying VLMs for high-level semantic exploration and interaction in embodied environments remains insufficient for solving such challenging tasks. Our benchmark, HomeSPT-Bench, is designed to expose these limitations and demonstrate that current VLM/LLM-based agents still struggle to perform robust proactive assistance in embodied environments. Even when these models correctly interpret dialogue instructions, they may still fail to ground the correct object, inspect relevant environmental states, avoid unnecessary actions, or terminate behavior appropriately once the task is completed.

The performance gap between dialogue-only settings and environment-grounded settings is particularly significant. This observation indicates that relying primarily on language priors, without sufficient integration of environmental feedback, is inadequate for real-world household scenarios. Instead, embodied task execution fundamentally requires tight coupling between language understanding and perceptual grounding in dynamic environments.

Motivated by these findings, HomeSPT-Bench emphasizes that proactive assistance should be evaluated not only by a model's

Table 2 Results across task categories on HomeSPT-Bench Unit: %

Task categories	Qwen3.7-Plus			GPT-5		
	Dialogue-only	Scene text	Active semantic	Dialogue-only	Scene text	Active semantic
Fetch object	7.5	12.1	55.4	8.8	13.3	57.1
State check	14.6	17.5	58.8	13.3	15.8	60.0
Environment adjustment	15.4	24.2	42.9	14.6	26.7	40.4
Rearrangement	5.4	14.2	32.1	6.2	13.3	34.6
Safety assistance	12.1	17.9	27.1	10.8	17.1	27.9
No action	7.9	15.8	26.2	7.1	17.5	24.6

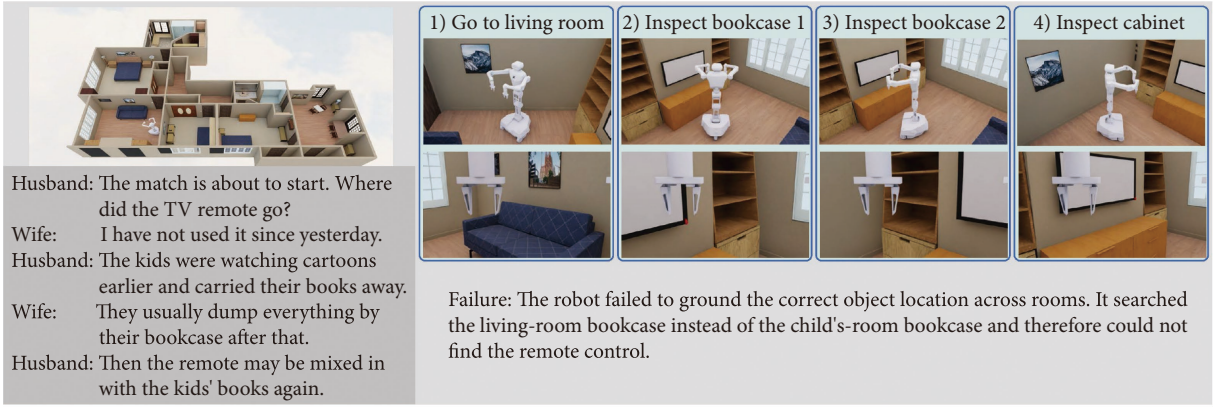


Figure 5 Failure case involving cross-room grounding, where the agent follows an incorrect room-level association and searches an irrelevant bookcase, missing the object needed for the task.

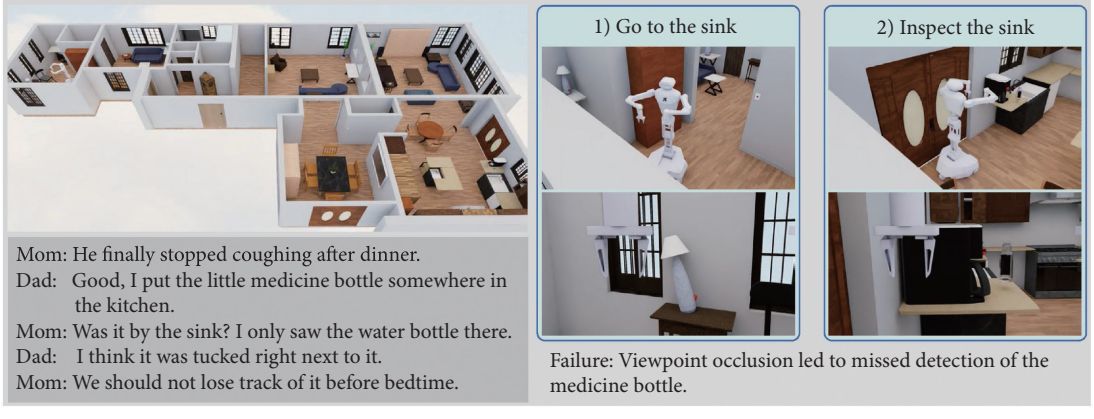


Figure 6 Failure case caused by insufficient visual evidence, where the target object is missed under occlusion, showing the need for stronger active perception and action selection.

ability to infer human intent from dialogue, but also by its capacity to ground that intent in the current physical scene and act accordingly.

5 Conclusions

This work studies a key yet insufficiently explored capability for embodied agents: the ability to recognize when everyday dialogue implies a latent need, determine whether intervention is appropriate, and ground the resulting decision in the current physical environment. Rather than treating proactive assistance as a direct extension of instruction following, we claim that it requires a distinct form of situated reasoning, where the appropriate response depends jointly on conversational context, object states, spatial layout, and the degree of assistance required by the situation.

To support research in this direction, we presented HomeSPT-Bench, a benchmark for situated proactive task generation in embodied environments. HomeSPT-Bench provides simulator-grounded household scenarios in which task-critical variables are fixed before dialogue synthesis, allowing implicit needs to be evaluated against concrete scene states and formally defined target outcomes. By covering multiple categories of proactive assistance, including no-action cases, the benchmark shifts the focus from executing pre-specified instructions to understanding when and how an embodied agent should proactively assist.

Our experiments reveal that current VLM-based agents still face substantial challenges in situated proactive assistance. Although additional scene information and active interaction provide useful environmental evidence, substantial limitations remain in spatial grounding, state verification, occlusion handling, and intervention decision-making. These findings indicate that robust proactive embodied behavior cannot be achieved through language reasoning alone. Instead, future agents will require more advanced and tightly integrated capabilities in perception, action, and reasoning, including active scene understanding, reliable environmental state tracking, context-sensitive decision-making, and adaptive interaction with the physical world.

HomeSPT-Bench is therefore intended not only as an evaluation resource, but also as a step toward reframing embodied assistance around appropriate, context-aware, and human-friendly proactivity. As embodied agents are deployed in increasingly complex human environments, reliable assistance will require capabilities beyond accurate command execution, including the ability to infer unspoken needs, calibrate an appropriate degree of initiative, and act in ways that are context-aware, socially appropriate, and aligned with human preferences.

Funding

This work was jointly supported by the National Natural Science Fundation of China (62120106005), and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM109).

Data availability

The data generated during this study are not publicly available at this time.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Author contribution statement

Haifeng Xu: Conceptualization, Methodology, Writing, Experiments. Chunwen Li: Investigation, Methodology. Xiaohu Yuan: Writing, Experiments. Huaping Liu: Conceptualization, Review & Editing, Supervision.

Use of AI statement

None.

References

[1] Baraglia J, Cakmak M, Nagai Y, et al. Efficient human-robot collaboration: when should a robot take initiative?[J]. *The International Journal of Robotics Research*, 2017, 36(5/7): 563-579.

[2] Patel M, Chernova S. Proactive robot assistance via spatio-temporal object modeling[C]//Proceedings of the 6th Conference on Robot Learning. Auckland: PMLR, 2023: 881-891.

[3] Peng Z H, Kwon Y, Lu J A, et al. Design and evaluation of service robot's proactivity in decision-making support process[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow: ACM, 2019: 98.

[4] Huang C M, Mutlu B. Anticipatory robot control for efficient human-robot collaboration[C]//2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Christchurch: IEEE, 2016: 83-90.

[5] Huang C M, Cakmak M, Mutlu B. Adaptive coordination strategies for human-robot handovers[C/OL]//Robotics: Science and Systems Conference, 2015:1-10. https://www.roboticsproceedings.org/rss11/p31.html.

[6] Koppula H S, Saxena A. Anticipating human activities using object affordances for reactive robotic response[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 14-29.

[7] Mascaro E V, Sliwowski D, Lee D. HOI4ABOT: human-object interaction anticipation for human intention reading collaborative roBOTs[C]//Proceedings of the 7th Conference on Robot Learning. Atlanta: PMLR, 2023: 1111-1130.

[8] Xu H F, Li Y C, Ma L M, et al. Autonomous embodied navigation task generation from natural language dialogues[J]. *Science China Information Sciences*, 2025, 68(5): 150208.

[9] Xu H F, Li C W, Yuan X H, et al. A simulated dataset for proactive robot task inference from streaming natural language dialogues[J]. *Scientific Data*, 2025, 12(1): 1405.

[10] Liu H P, Guo D, Cangelosi A. Embodied intelligence: a synergy of morphology, action, perception and learning[J]. *ACM Computing Surveys*, 2025, 57(7): 186.

[11] Kolve E, Mottaghi R, Han W, et al. AI2-THOR: an interactive 3D environment for visual AI[PP/OL]. arXiv (2017-12-14)[2026-06-01]. https://arxiv.org/abs/1712.05474. arXiv:1712.05474.

[12] Savva M, Kadian A, Maksymets O, et al. Habitat: a platform for embodied AI research[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9338-9346.

[13] Puig X, Undersander E, Szot A, et al. Habitat 3.0: a co-habitat for humans, avatars, and robots[C]//The 12th International Conference on Learning Representations. Vienna: OpenReview. net, 2024: 15306-15336.

[14] Puig X, Ra K, Boben M, et al. VirtualHome: simulating household activities via programs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8494-8502.

[15] Deitke M, Vander Bilt E, Herrasti A, et al. ProcTHOR: large-scale embodied AI using procedural generation[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc. , 2022: 433.

[16] Li C S, Zhang R H, Wong J, et al. BEHAVIOR-1K: a benchmark for embodied AI with 1, 000 everyday activities and realistic simulation[C]//Proceedings of the 6th Conference on Robot Learning. Auckland: PMLR, 2023: 80-93.

[17] Shridhar M, Thomason J, Gordon D, et al. ALFRED: a benchmark for interpreting grounded instructions for everyday tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10737-10746.

[18] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3674-3683.

[19] Hong Y C, Wu Q, Qi Y K, et al. VLN BERT: a recurrent vision-and-language BERT for navigation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1643-1653.

[20] Das A, Datta S, Gkioxari G, et al. Embodied question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1-10.

[21] Tan S N, Ge M M, Guo D, et al. Knowledge-based embodied question answering[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(10): 11948-11960.

[22] Padmakumar A, Thomason J, Shrivastava A, et al. TEACh: task-driven embodied agents that chat[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI Press, 2022: 2017-2025.

[23] Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: grounding language in robotic affordances[C]//Proceedings of the 6th Conference on Robot Learning. Auckland: PMLR, 2023: 287-318.

[24] Li X H, Li P Y, Qian L, et al. What matters in building vision–language–action models for generalist robots[J]. *Nature Machine Intelligence*, 2026, 8(2): 158-172.

[25] Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model[C]//Proceedings of the 8th Conference on Robot Learning. Munich: PMLR, 2025: 2679-2713.

[26] Li X H, Liu M H, Zhang H B, et al. Vision-language foundation models as effective robot imitators[C]//The 12th International Conference on Learning Representations. Vienna: OpenReview. net, 2024: 26703-26721.

[27] Cao Z H, Wang Z D, Xie S W, et al. Smart help: strategic opponent modeling for proactive and adaptive robot assistance in households[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 18091-18101.

[28] Tanneberg D, Ocker F, Hasler S, et al. To help or not to help: LLM-based attentive support for human-robot group interactions[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi: IEEE, 2024: 9130-9137.

[29] Grannen J, Karamcheti S, Wulfe B, et al. ProVox: personalization and proactive planning for situated human-robot collaboration[J]. *IEEE Robotics and Automation Letters*, 2025, 10(8): 8451-8458.

[30] Lu Y X, Yang S Z, Qian C, et al. Proactive agent: shifting LLM agents from reactive responses to active assistance[C]//The 13th International Conference on Learning Representations. Singapore: OpenReview. net, 2025: 47431-47457.

[31] Wang S Y, Fu J L, Liu F H, et al. RoboOmni: proactive robot manipulation in omni-modal context[C/OL]//The 14th International Conference on Learning Representations, 2026: 30282-30314. https://proceedings.iclr.cc/paper_files/paper/2026/hash/33610fba262d7b6fed0810b89f55e147-Abstract-Conference.html.