ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust Robotic Task Planning via Immutable Subgoals

Chulyong Lim
Chung-Ang University
dlacjfdyd@cau.ac.kr

Jaewon Baek
Chung-Ang University
bjw4945@cau.ac.kr

Junhee Han
Chung-Ang University
han8760@cau.ac.kr

Wooyeol Bae
Chung-Ang University
baewy5612@cau.ac.kr

Woochul Nam*
Chung-Ang University
wcnam@cau.ac.kr

***Abstract* - Service robots require instruction-following capabilities to perform various tasks regardless of environmental changes. A task planner must accurately infer user intent even when human instructions are ambiguous. To this end, we propose TIGER, a task planning framework that generates reliable action sequences by deriving immutable subgoals from instructions. TIGER employs an Immutable Subgoal Planner (ISP) to decompose instructions into environment-independent subgoals and a Target Grounder (TG) to ground abstract keywords to real-world objects via visual perception and reasoning. A task-representative one-shot strategy improves subgoal generation using only seven annotated examples. TIGER outperformed LLM-Planner in the ALFRED benchmark, increasing success rates from 15.09% to 35.06% on the seen set and from 19.73% to 42.57% on the unseen set. Its scalability was also verified in real-world experiments with a UR5e robot.**

***Index Terms* - AI-Enabled Robotics, Instruction following, Large Language Model, Task and Motion Planning, Visual grounding, Vision-Language Model**

INTRODUCTION

The capability of Embodied Instruction Following has emerged as a key technology for enabling robots to adapt to diverse environments and accomplish complex tasks. Rapid advancements in large language models (LLMs) [1] have opened new possibilities for robots to understand natural language instructions and generate complex task plans through few-shot learning. Despite these advancements, LLM-based task planning approaches still face fundamental limitations. Prior studies [2] have identified three common failure modes: (1) dependency violations, (2) object hallucinations, and (3) syntactic errors.

To address these limitations, this study proposes TIGER, a new task planning framework with four key components:

• Keyword Extractor (KE): extracts task-relevant keywords from natural language instructions.

• Immutable Subgoal Planner (ISP): decomposes instructions into environment-independent immutable subgoals, reducing the effective planning horizon.

• One-shot strategy: plans each task type using a single representative example, requiring only seven annotations.

• Target Grounder (TG): grounds abstract keywords to real-world objects via visual perception and chain-of-thought reasoning.

The main contributions are: (1) ISP that enables robust planning for long-horizon tasks, (2) a one-shot strategy that reduces annotation effort while maintaining competitive performance, (3) TG, a multi-stage visual grounding pipeline for cluttered scenes, and (4) comprehensive validation across simulation and real-world environments.

PROPOSED METHOD

TIGER extracts task-relevant keywords from instructions using KE, then ISP generates a sequence of immutable subgoals regardless of environmental conditions. Each subgoal keyword is grounded to a real-world object through TG. The low-level policy module is interchangeable, allowing operation across simulation and real-world platforms.

The Immutable Subgoal Planner overcomes the hallucination problem of previous LLM-based planners that directly generate long-horizon primitive action

sequences. Instead of generating low-level actions directly, it produces concise immutable subgoals such as Find(egg), Heat(egg), and Put(egg, sink), each of which is assigned to a corresponding low-level policy. Because these subgoals are defined at the semantic level, they remain valid regardless of the environment state.

Whereas LLM-Planner [3] requires 100 annotated examples, the task-representative one-shot strategy in TIGER selects a single example per task type and uses only seven examples to cover all ALFRED task types. This reduces annotation effort by 14 times while enabling more accurate plan generation.

The Target Grounder selects appropriate objects through three stages. The Shopper first ranks candidate objects by relevance using a VLM. The Target-Attention Model then localizes the chosen candidate using Grounding DINO [4] and produces a high-resolution crop. Finally, the Judge verifies appropriateness through chain-of-thought prompting before the object is grounded.

RESULTS

ALFRED Benchmark. TIGER was evaluated on the ALFRED benchmark [5], which includes seven housework task types. The same low-level policy was applied to all models for fair comparison. With only seven annotated examples, TIGER with GPT-4o achieved 35.06% SR (seen) and 42.57% (unseen), far exceeding LLM-Planner with GPT-4o (15.09% / 19.73%) that uses 100 examples. TIGER with GPT-4o-mini also outperformed LLM-Planner with GPT-4o, demonstrating that ISP extracts better plans regardless of the underlying LLM.

Real-World Experiments. A UR5e robot was controlled by TIGER in a multi-room environment with four tasks of varying instruction ambiguity. TIGER achieved 76.7%–100.0% SR, while ProgPrompt [6] and Inner Monologue [7] showed 0.0%–43.3%. TG effectively identified appropriate objects even in cluttered scenes.

CONCLUSION

TIGER addresses long-horizon hallucination problems of LLM-based planners by introducing immutable subgoals through ISP, enabling geometrically feasible task execution even in unseen environments. TG improves object grounding in cluttered scenes, and the one-shot strategy maintains competitive performance with far less annotation than conventional approaches. Future work will extend TIGER along three directions: a failure-handling module that provides execution feedback for automatic re-planning, more robust target grounding under visual confusion in cluttered scenes, and noise-tolerant low-level policies for reliable subgoal execution in real-world environments.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-02214162). This work was also supported by the National Research Foundation of Korea (NRF) under the BK21 FOUR program (Intelligent Wearable Education Research Center).

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[2] B. Kim, J. Kim, Y. Kim, C. Min, and J. Choi, “Context-aware planning and environment-aware memory for instruction following embodied agents,” in Proc. IEEE/CVF ICCV, pp. 10936-10946, 2023.

[3] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, “LLM-Planner: Few-shot grounded planning for embodied agents with large language models,” in Proc. IEEE/CVF ICCV, pp. 2998-3009, 2023.

[4] S. Liu, Z. Zeng, T. Ren, et al., “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” in ECCV, pp. 38-55, 2024.

[5] M. Shridhar, J. Thomason, D. Gordon, et al., “ALFRED: A benchmark for interpreting grounded instructions for everyday tasks,” in Proc. IEEE/CVF CVPR, pp. 10740-10749, 2020.

[6] I. Singh, V. Blukis, A. Mousavian, et al., “ProgPrompt: Generating situated robot task plans using large language models,” in IEEE ICRA, pp. 11523-11530, 2023.

[7] W. Huang, F. Xia, T. Xiao, et al., “Inner Monologue: Embodied reasoning through planning with language models,” in CoRL, pp. 1769-1782, 2022.