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

GPT-PDDL: Towards Executable Robot Task Planning

Changsik Lee
Human-Centric Robotics R&D Dept.
Korea Institute of Industrial Technology
Asan 15588, Rep. of Korea
cslee@kitech.re.kr

Hye-Kyung Cho
Deprartment of Applied AI
Hansung University
Seoul 02876, Rep. of Korea
hkcho@hansung.ac.kr

Sujeong You
Human-Centric Robotics R&D Dept.
Korea Institute of Industrial Technology
Asan 15588, Rep. of Korea
sjyou21@kitech.re.kr

***Abstract*— Given the recent significant advancements in the video understanding capabilities of Large Language Models (LLMs), there is growing interest in research that automatically generates executable robot task plans from human demonstration videos. Existing LLM-based symbolic planning approaches often rely on manually defined Problem Domain Definition Language (PDDL) domains or fixed action primitives. This paper proposes GPT-PDDL, a framework that infers step-by-step task procedures from demonstration videos and converts them into robot plans based on PDDL.**

Keywords*— *LLM, PDDL, task and motion planning, robot action, human demonstration

I. INTRODUCTION

For a robot to perform tasks at a human level, the ability to understand complex sequences of task procedures and convert them into execution plans is essential. Traditionally, task plans were primarily defined manually by humans or extracted through sensor-based demonstration tracking. However, these methods are time-consuming and difficult to generalize to new tasks.

Large Language Models (LLMs) possess high-level semantic understanding capabilities-such as video summarization, action description generation, and causal reasoning-making them suitable tools for solving the problem of plan generation based on demonstration videos. A recent study[1] proposed a one-shot visual teaching pipeline that utilizes a general-purpose vision-language model, GPT-4V, to generate executable robot task plans by observing human demonstration videos, as shown in Fig. 1. While this pipeline can be applied to various scenarios through prompt engineering, LLM hallucinations can result in erroneous object recognition or unfeasible action plans. Although some methods utilize prior knowledge [2] to verify task plans generated by language models, they are still insufficient in guaranteeing the feasibility of task execution.

To mitigate these hallucinations, integrating a PDDL Planner enables the generation of robot task plans that can automatically verify physical feasibility[3]. Although PDDL has been employed for logical verification, there has been limited investigation into its integration with VLM-based pipelines for inferring robot task plans from human video demonstrations

The goal of this research is to integrate the video understanding capabilities of LLMs with a traditional planning language (PDDL) to build a system where a robot can automatically obtain a structured task plan simply from a human demonstration.

II. SYSTEM ARCHITECTURE AND METHODOLOGY

GPT-PDDL consists of three stages: (1) video-text extraction, (2) action unit segmentation and intent inference, and (3) PDDL plan conversion. Demonstration videos are processed frame by frame, and GPT-series models generate detailed action descriptions that reflect temporal context.

The model generates mid-level actions based on language, such as "pick," "place," "rotate," and "open," rather than low-level primitive actions. This reduces the reasoning gap between natural language and planning, making it easier to map to common robot task domains. Subsequently, the system constructs a set of PDDL fluent candidates by inferring the sequential relationships, preconditions, and effects between actions.

The inferred action structure is then automatically converted into PDDL actions by combining it with predefined domain templates. Using GPT, the system maps actions described in natural language to the PDDL action schema, performing automatic object type classification and state variable estimation when necessary. Finally, a classical planner (e.g., FastDownward) is used to verify executability and generate the final plan.

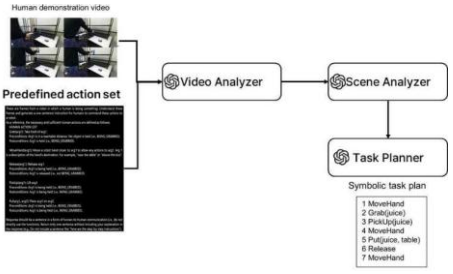

Fig. 1. the baseline[1] framework.

III. RESULTS

To verify the effectiveness of the proposed method, the RH20T[4] dataset was utilized. The RH20T[4] dataset

includes paired human and robot task videos for the same tasks and consists of a total of 147 tasks. In this paper, five specific tasks were selected for the experiment and provided as input to both the proposed framework and the baseline framework.

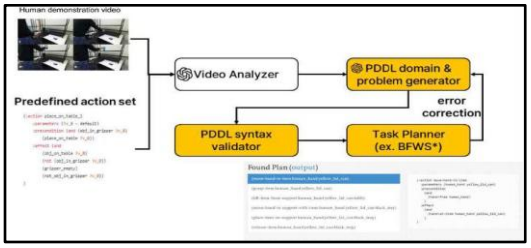

Fig. 2. Proposed GPT-PDDL pipeline.

Experimental details are as follows:

- **Action Sequence**: The robot's action sequence is planned by inputting task demonstration videos and task instructions, along with prompts, into a Vision-Language Model (VLM). can stand alone in conveying the results..
- **Verification**: Verifying the logical consistency of task plans generated by VLM through PDDL (Planning Domain Definition Language) representation methods.
- **Task Planning**: Generating robot action sequences using the PDDL solver POPF (Partial Order Planning Forwards).
- **Experimental Data**: A total of 20 demonstration videos of human tasks extracted from the RH20T[4] dataset

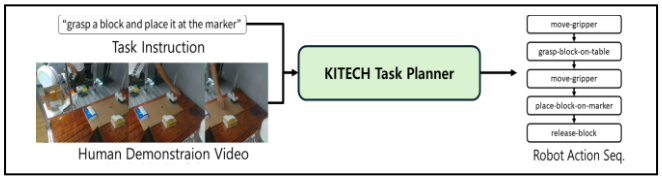

Fig. 3. Overvall pipeline

TABLE I. SELECTED VIDEOS AND CHANGES

No.	Video Content	Changes
1		• Original Video
2		• Object Location
3		• Object Location • Camera View
4		• Object Location
5		• Object Location • Camera View • Object Size
6		• Object Location • Camera View • Object Size
7		• Object Location • Camera View • Object Size
8		• Another Subejct • Object Location • Table • Clutter Background

TABLE II. PERFORMANCE COMPARISON OF 5 TASKS ON THE RH20T[4] DATASET

Task no.	Mild	Baseline[1]	Ours
8	grab a block and place it at the designated location	90%	100%
13	place the block on the scale	90%	100%
14	remove the object from the scale	90%	100%
34	stack the squares into a pyramidic shape	60%	70%
37	stack the blocks in a vertical line of five	60%	80%
	Average Success Rate	78%	**90%**

IV. CONCLUSIONS

Traditional LLM-based planning often suffers from hallucinations or physically impossible steps. We address this by using PDDL (Planning Domain Definition Language) as a formal intermediary. This paper proposes the GPT-PDDL method, which combines GPT-based video understanding with PDDL-based plan generation. By converting human demonstration directly into a structured planning language, the proposed approach has the advantage of understanding high-level task structures and converting them into executable plans using only demonstration videos. This pipeline allows robots to perform complex tasks with high reliability.

ACKNOWLEDGMENT

This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (Project Number: P0028922).

REFERENCES

[1] Wake N., Kanehira A., Sasabuchi K., Takamatsu J., and Ikeuchi K. "GPT-4V(ision) for Robotics: Multimodal Task Planning from Human Demonstration," in IEEE Robotics and Automation Letters, pp. 10567-10574, vol. 9, no. 11, Oct. 2024.

[2] Chen G. et al., "Human Demonstrations are Generalizable Knowledge for Robots," lEEE/RSJ international Conference on Intelligent Robots and Systems (IROS), Oct. 2025.

[3] Guan L., Valmeekam K., Sreedharan S., and Kambhampati S. Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning," 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Dec. 2023.

[4] H.-S. Fang et al., "RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot," IEEE International Conference on Robotics and Automation (ICRA), 2024.