MANCHESTER 1824
The University of Manchester

GPT-4V(ision) for Robotics: Multimodal Task Planning from Human Demonstration

Applied Robotics Research, Microsoft

Yang Li
PhD Student of Computer Science
University of Manchester

The University of Manchester

Outlines

- **Introduction**
  - **Overview**
  - **Related Work**
  - **Contributions**
- **Method**
  - **Overview**
  - **Symbolic Task Planner**
  - **Affordance Analyzer**
- **Experiments**
- **Conclusion**
- **Discussion**

MANCHESTER 1824
The University of Manchester

Introduction - Overview

The paper presents a novel ***pipeline*** using the ***multimodal models GPT-4V(ision) and GPT-4***, that enables learning ***robotic manipulation*** from ***human action observations (video) and text instruction***.

Key Insight - Why use off-the-shelf models like GPT?

One key advantage of using off-the-shelf models is their flexibility; they can be adapted to various robotic hardware configurations and functionalities simply by modifying prompts. This approach removes the necessity for extensive data collection and model retraining for different hardware or scenarios, greatly improving system reusability in research and easing the transition to industrial applications.

Hence, utilizing off-the-shelf models for robot manipulation represents a promising direction.

MANCHESTER 1824
The University of Manchester

Introduction – Related Work

1. LLM/VLM-based task planning

End-to-End Training using specific dataset:

- **For example:** Brohan et al. proposed a transformer-based model that trained based on both robotic trajectory data and internet-scale vision-language tasks
- **Disadvantages:** 1. Require a large amount of data 2. necessitate data recollection and model retraining when transferring or extending these to other robotic settings

Utilizing off-the-shelf LLMs:

- Decomposing human instructions into high-level subgoals, while pre-trained skills achieve the subgoals.
- is typically seen as a part of framework, called task and motion planning (TAMP)

This Work ⟹ **Extending the method to multi-modal input by leveraging off-the-shelf GPT-4V and GPT-4**

Introduction – Related Work

2. Grounding visual information for robotics

LLMs-based task planning: Executing long task steps as planned is often challenging due to **unforeseen and unpredicted environmental situations**

- Current strategies seek to achieve environment grounded robot execution by integrating **environmental information and adjusting the robot's actions** at the task plan or controller level.

This Work ⇒ **Using GPT-4V to ground environment information**: open-vocabulary object detector and relationship between the hand and the object

3. Learning Affordance

Affordance: refers to the **potential for action** that objects or situations in an environment **provide to an individual** [1]

- **In the field of robotics**, it often pertains to the meaning of executable actions in that environment, and information about areas **where action is possible**.

[1]. J. J. Gibson, The ecological approach to visual perception: classic edition. Psychology press, 2014.

Introduction – Related Work

3. Learning Affordance

Affordance: refers to the **potential for action** that objects or situations in an environment **provide to an individual**

Current Methods more focus on feasibility of planed task

- calculates the **feasibility** of robotic functions from **visual information and compares it with planned tasks** [1]
- using LLMs/VLMs to extract the knowledge of movable area [2]

⇩

Focusing on the **relationship between the working environment, the objects being manipulated, and the robot**, **Affordance** can be considered that object manipulation involves even more constraints.

- Like: waypoints for collision avoidance, grasp types, and upper-limb postures

⇨ **Those information is often not taught explicitly. Thus, this work proposes a pipeline to extract those affordance information.**

[1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al., "Do as i can, not as i say: Grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.
[2] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," arXiv preprint arXiv:2307.05973, 2023.

MANCHESTER 1824
The University of Manchester

Introduction – Contributions

1. Proposing a ready-to-use multimodal task planner that utilizes off-the shelf VLM and LLM
2. Proposing a methodology for aligning GPT-4V's recognition with affordance information for grounded robotic manipulation
3. Making the code publicly accessible as a practical resource for the robotics research community

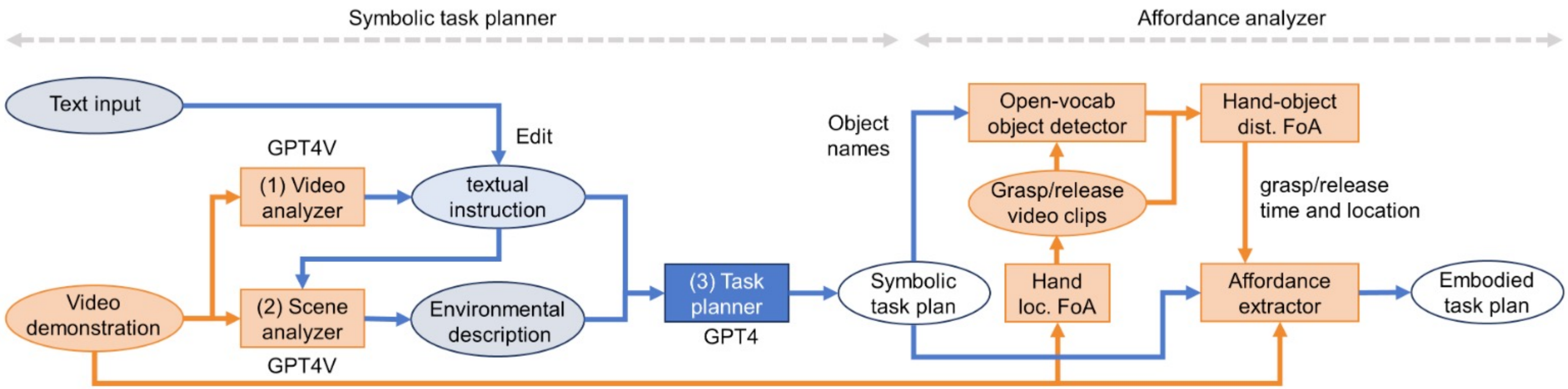
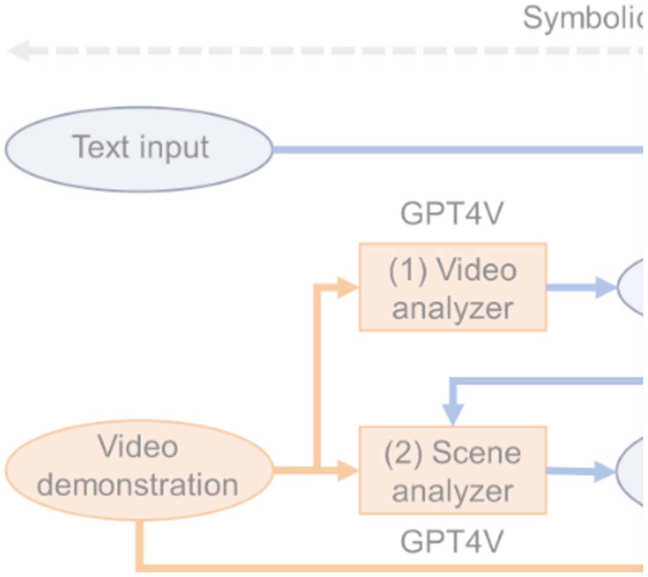


Fig. 2. Proposed pipeline of the multimodal task planner. It consists of the symbolic task planner and the affordance analyzer. Blue components/lines are text-based information, and the red components are vision-related information. FoA denotes focus-of-attention.

Introduction – Contributions

1. Proposing a ready-to-use m … l LLM
2. Proposing a methodology … nformation for grounded robotic manipula…
3. Making the code publicly a … community



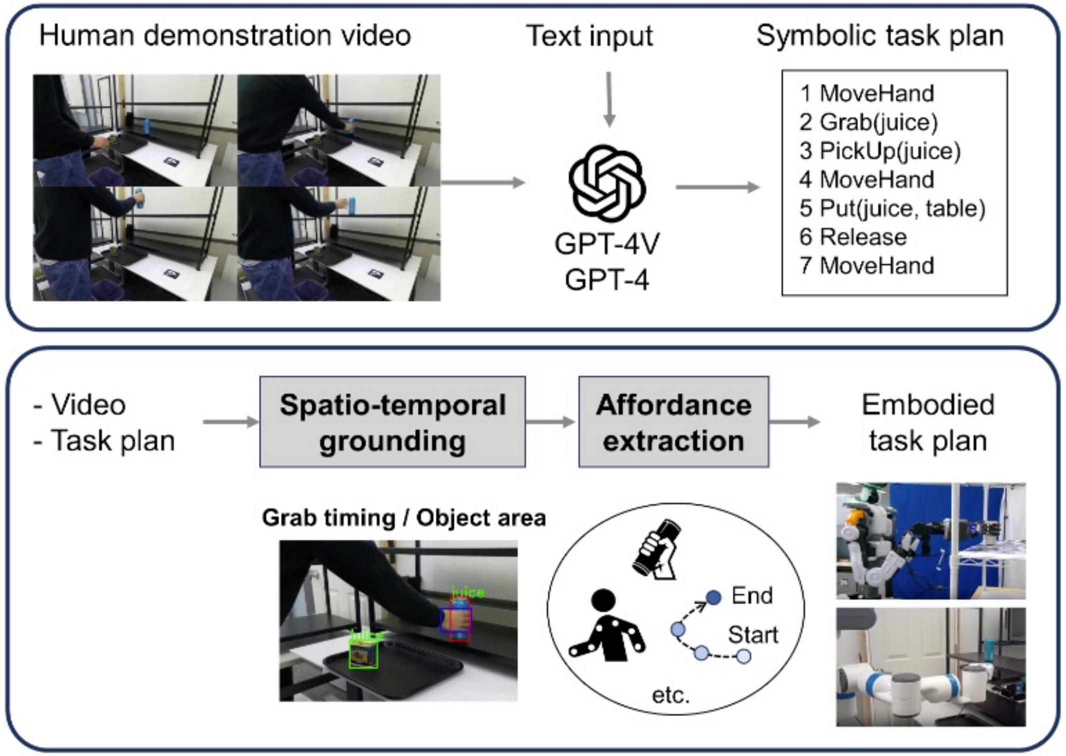


Fig. 1. This figure illustrates the proposed multimodal task planner utilizing GPT-4V and GPT-4. It highlights the system's ability to process video demonstrations and text instructions, generating task plans and extracting key affordances for robotic execution, which are then compiled into a JSON format.

Fig. 2. Proposed pipeline of the multimod … nce analyzer. Blue components/lines are text-based information, and the red compon…

Method - Overview

The proposed system is composed of two pipelines connected in series

- **Symbolic task planner**: takes teaching videos, text, or both as input, then outputs a sequence of robot actions.
- **Affordance analyzer:** analyzes the video to determine when and where the tasks occur, and then extracts the affordance information necessary for efficient task executions.

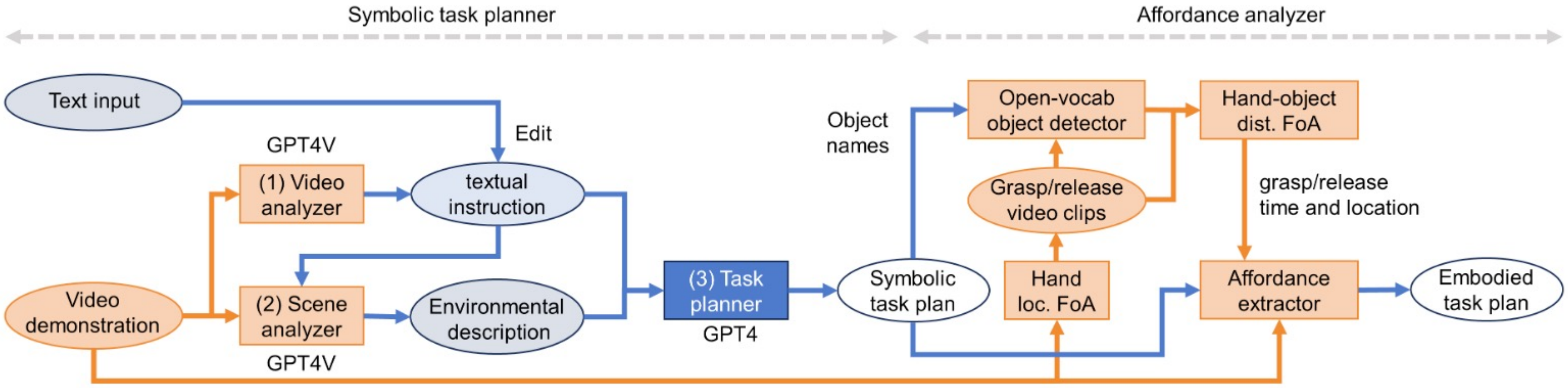


Fig. 2. Proposed pipeline of the multimodal task planner. It consists of the symbolic task planner and the affordance analyzer. Blue components/lines are text-based information, and the red components are vision-related information. FoA denotes focus-of-attention.

The input videos are **demonstrations** of humans performing actions that **are intended to be replicated by the robot**.

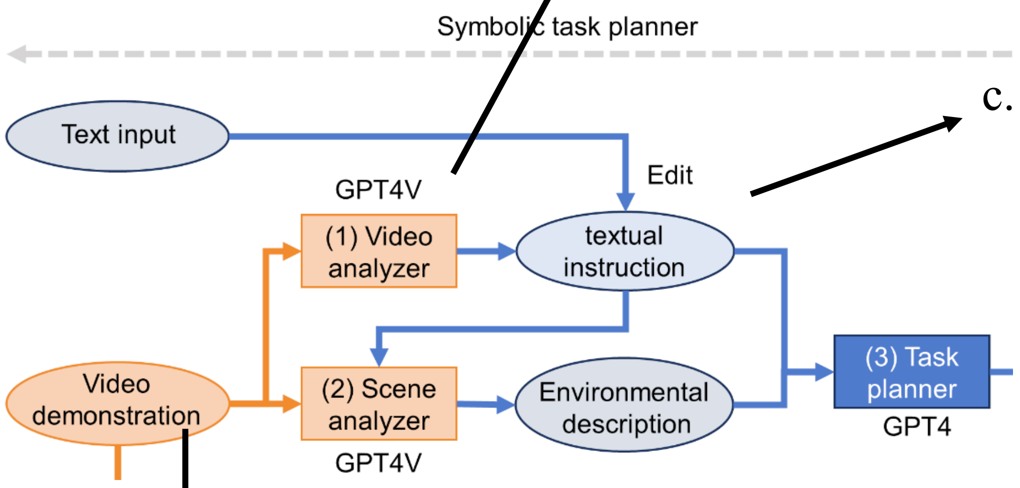
Method – Symbolic Task Planner

Symbolic task planner: takes teaching videos, text, or both as input, then outputs a sequence of robot actions.

1. video analysis

b. transcribes them into text instructions in a style used in human-to-human communication (for example, 'Please throw away this empty can').

c. The output text is then checked and edited by the user



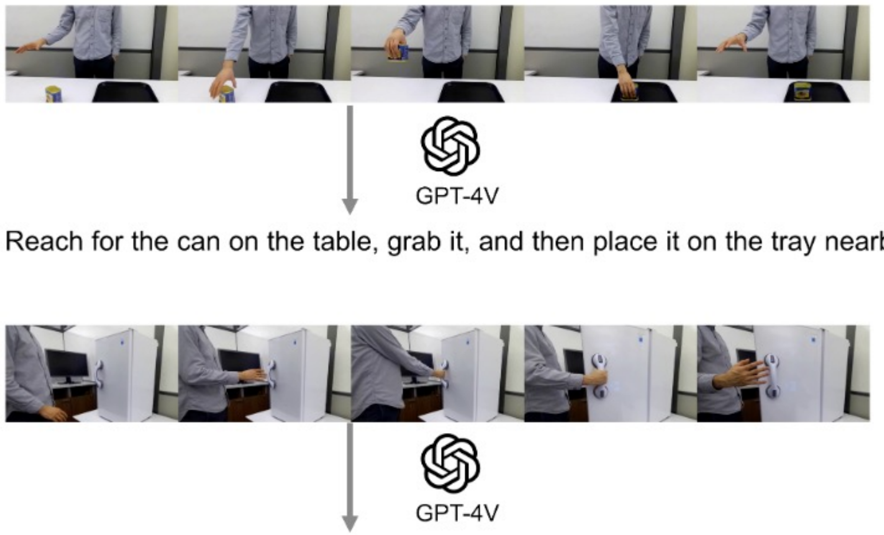
a. frames are extracted at regular intervals rather than from every frame and fed into GPT-4V.

Metho… …er

Symbolic task planner: takes tea… …equence of robot actions.

1. video analysis,

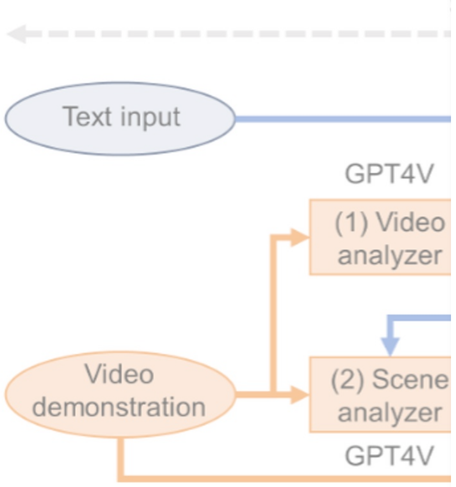
These are frames from a video in which a human is doing something. Understand these frames and generate a one-sentence instruction for humans to command these actions to a robot. As a reference, the necessary and sufficient human actions are defined as follows:
<…action list information…>
Response should be a sentence in a form of human-to-human communication (i.e., do not directly use the functions). Return only one sentence without including your explanation in the response.



Please move your hand to the fridge handle, grab it, and open the fridge door.

Fig. 3. Output of the video analyzer. Top pane shows the prompt for the GPT-4V and the bottom pane shows the examples of its output for two representative videos. The five frames are extracted at regular intervals and fed into GPT-4V. Part of the prompt is shown and the whole prompt is available at https://microsoft.github.io/GPT4Vision-Robot-Manipulation-Prompts/

…used in human-to-human
…this empty can').



…hecked and edited by the user

a. frames are extracted at re… …into GPT-4V.

Method – Symbolic Task Planner

Symbolic task planner: takes teaching videos, text, or both as input, then outputs a sequence of robot actions.

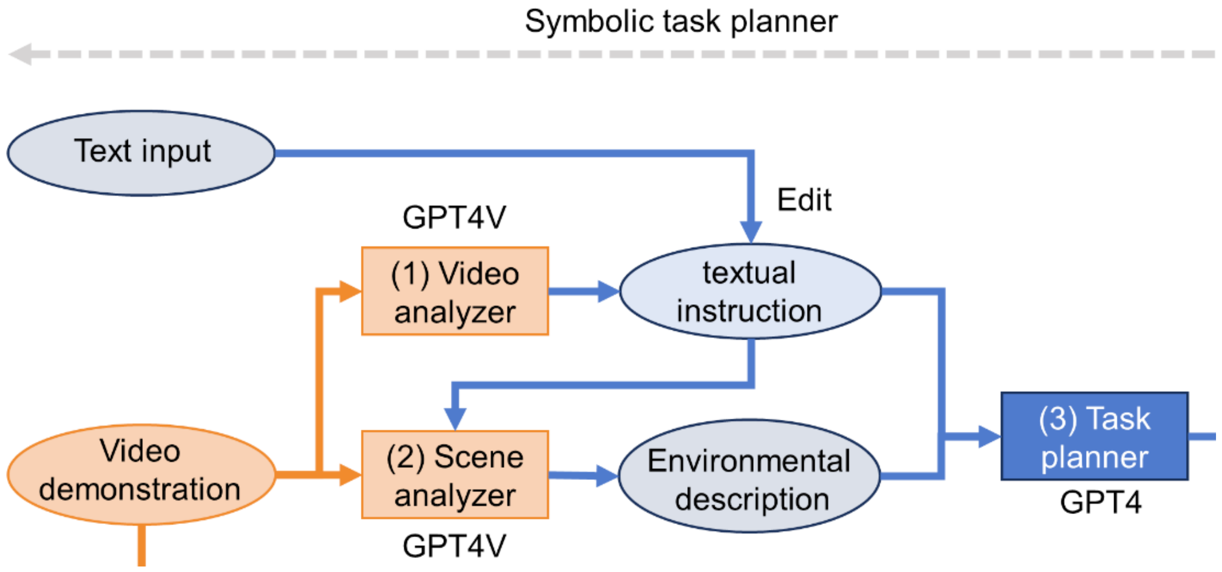
1. video analysis → 2. Scene Analysis

the first frame of the video data
or an image of the work environment
and the instructions

Scene Analyzer

Environmental Description

- a list of object names recognized by GPT-4V
- the graspable properties of objects
- the spatial relationships between objects



Besides, they prompted GPT-4V to explain the results of the object selection process and the reasons behind those choices.

MANCHESTER 1824
The University of Manchester

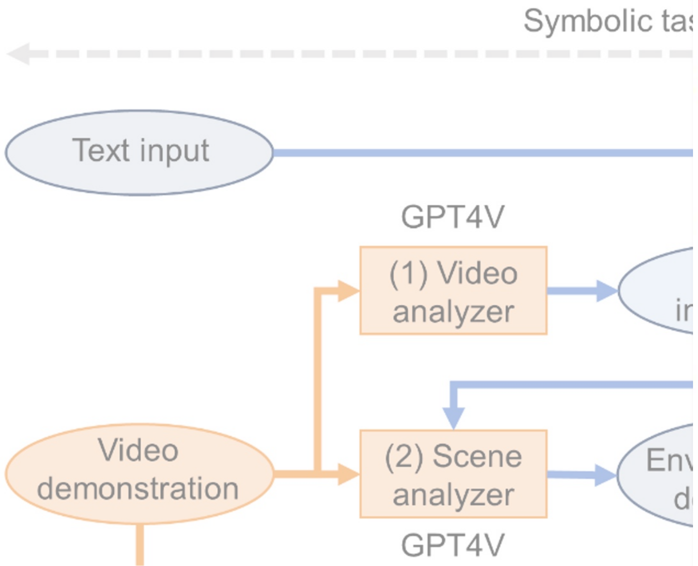
Method

Symbolic task planner: takes teach … nce of robot actions.

1. video analysis → 2. Scene An…

the first frame of the video data
or an image of the work environmen…
and the instructions

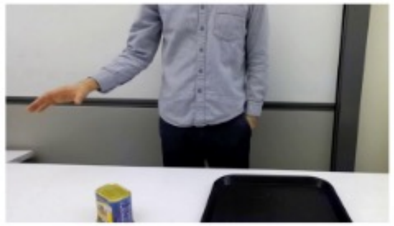
… ect names recognized by GPT-4V
… le properties of objects
… elationships between objects



… ompted GPT-4V to explain the
… ject selection process and the
… those choices.

Generated Instruction

This is a scene in which a robot is going to do "[ACTION]". Understand this scene and generate a scenery description to assist in task planning: Information about environments is given as python dictionary. For example:
<…example and formatting rules…>
Please take note of the following.
1. Focus only on the objects related to the given action information, if any.
2. The response should be a Python dictionary only, without any explanatory.



GPT-4V

GPT-4V

```
"objects": ["<can>", "<table>", "<tray>"],
"object_properties": {
    "<can>": ["GRABBABLE"],
    "<table>": [],
    "<tray>": []
},
"spatial_relations": {
    "<can>": ["on(<table>)"],
    "<table>": [],
    "<tray>": ["on(<table>)"]
},
"your_explanation": "The human is expected to grab the can that is on the table and place it on the tray. The table and the tray are included as they are the surfaces involved in this action. The can is marked as GRABBABLE because it is the item being manipulated. Other objects and details in the environment not related to the action, such as the human, the background elements, and the clothing, are omitted because they do not influence the task of moving the can from table to tray."
```

```
"objects": ["<fridge>", "<fridge_handle>"],
"object_properties": {
    "<fridge>": [],
    "<fridge_handle>": ["GRABBABLE"]
},
"spatial_relations": {
    "<fridge_handle>": ["on(<fridge>)"],
    "<fridge>": []
},
"your_explanation": "Human is moving their hand to the fridge handle, grabbing it, and opening the fridge door. Objects included are the fridge and fridge handle, as they are directly involved in the action. I omitted the computer monitor and other unmanipulated objects in the setting because they are not being interacted with for this task."
```

Yang Li (yang.li-4@manchester.ac.uk)

Method – Symbolic Task Planner

Symbolic task planner: takes teaching videos, text, or both as input, then outputs a sequence of robot actions.

1. video analysis → 2. Scene Analysis →3. Task Planning

Text Instructions
&
Environmental information

a prompt to make GPT-4 decompose given instruction into a sequence of robot tasks

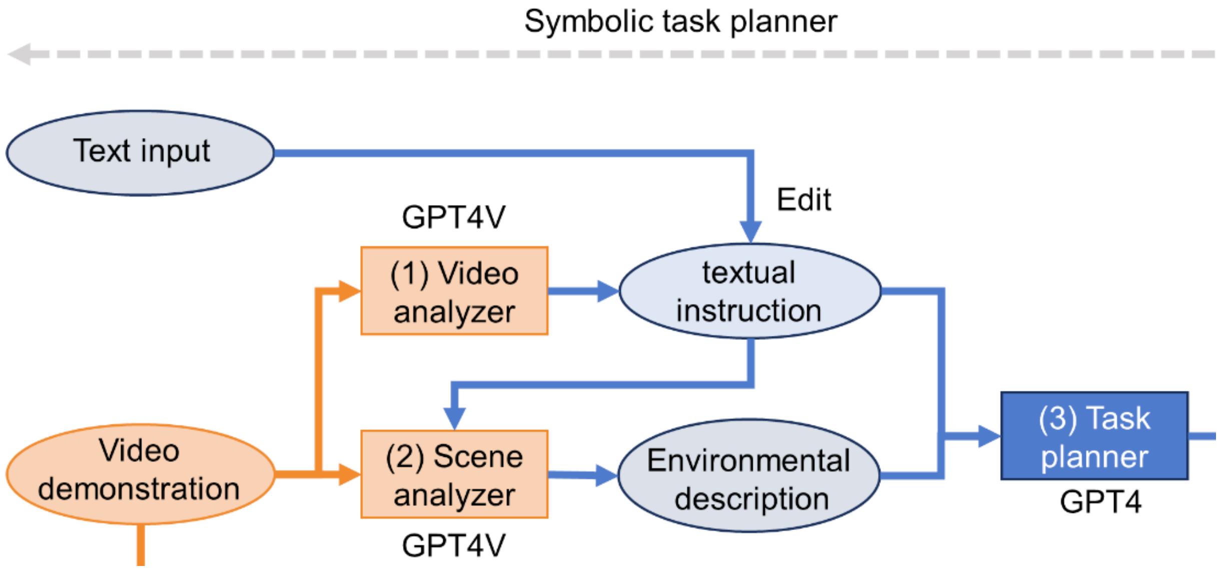


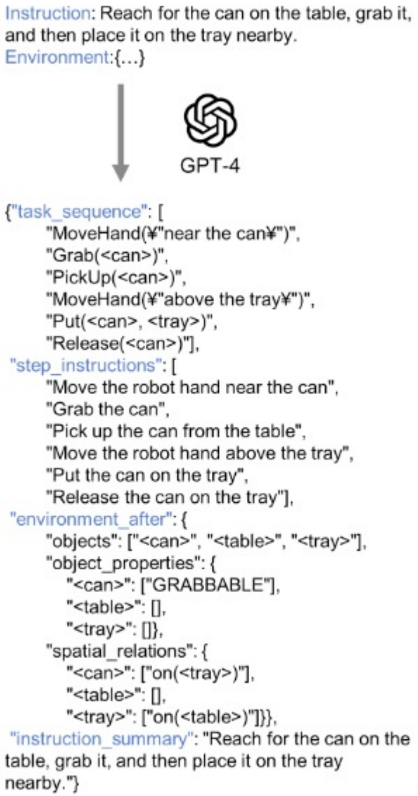
TABLE I
ROBOTIC ACTIONS AND THEIR EXPLANATIONS

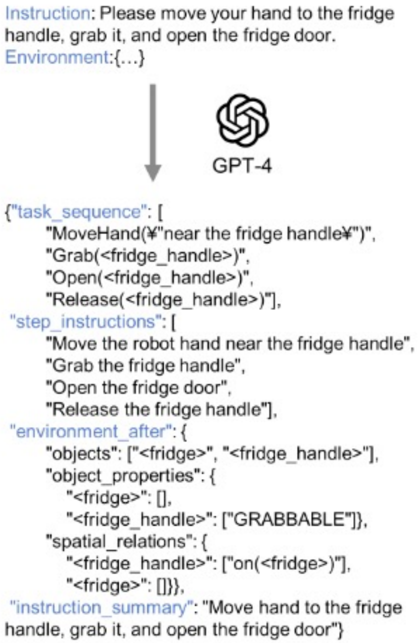
Action	Description
Grab(arg1)	Take hold of arg1. Preconditions: Arg1 is within reachable distance and no object is currently held. Postconditions: Arg1 is being held.
MoveHand(arg1)	Move the robot hand closer to arg1, where arg1 describes the hand's destination.
Release(arg1)	Release arg1. Preconditions: Arg1 is currently being held. Postconditions: Arg1 is no longer held.
PickUp(arg1)	Lift arg1. Preconditions: Arg1 is currently being held. Postconditions: Arg1 continues to be held.
Put(arg1, arg2)	Place arg1 onto arg2. Preconditions: Arg1 is currently being held. Postconditions: Arg1 continues to be held.
Rotate(arg1)	Open or close something by rotating arg1 along an axis. Preconditions: Arg1 is currently being held. Postconditions: Arg1 continues to be held.
Slide(arg1)	Open or close something by linearly moving arg1 along an axis. Preconditions: Arg1 is currently being held. Postconditions: Arg1 continues to be held.
MoveOnSurface(arg1)	Move arg1 across a surface. Preconditions: Arg1 is currently being held. Postconditions: Arg1 continues to be held.

Method – Symbolic Task Planner

Symbolic task planner: takes teaching videos, text, or both as input, then outputs a sequence of robot actions.

1. video analysis → 2. Scene Analysis →3. Task Planning





Output:

- task sequence
- output explanations for the tasks,
- estimated environments after the operation
- the summary of the operation,
- a set of task plans

Additionally, the task planner is a stateful system that **maintains a history of past conversations** within the token limits of the GPT-4 model.

Users can **modify** and **confirm** the output through linguistic feedback based on the output of the task planner.

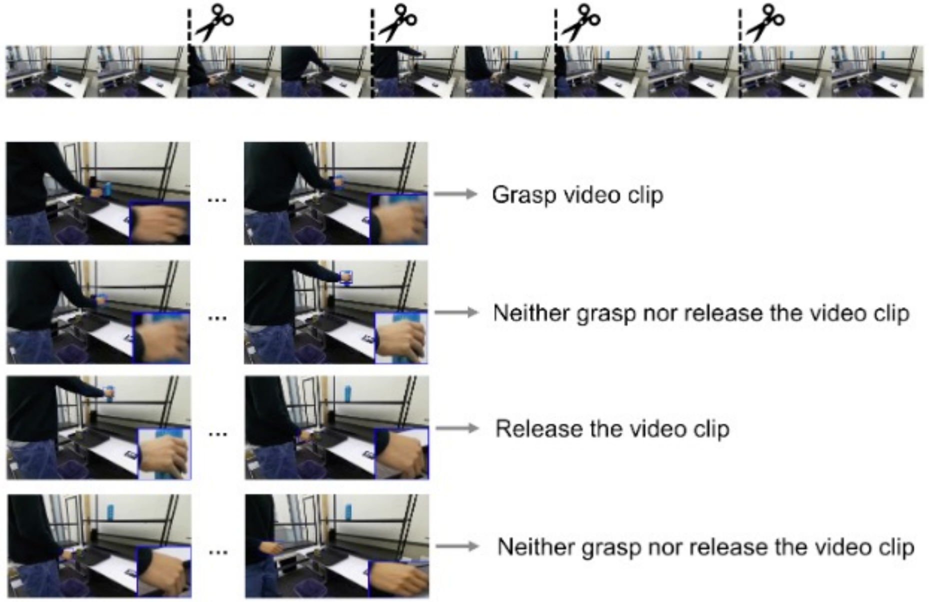
The University of Manchester

Method – Affordance Analyzer

Affordance Analyzer: **reanalyzes** the given videos using the knowledge from the symbolic task planner to acquire the **affordance information necessary** for the **robot's effective execution**.
Specifically, it focuses on the relationship between hands and objects based on the task's nature and object names.

- **Attention to Human Hands to Detect Grabbing and Releasing**

The pipeline divides a series of videos into video clips at regular time intervals.
The beginning and end frames of each video clip are then analyzed using a hand detector (YOLO-based) and an image classifier that determines whether an object is being grasped or not.



Method – Affordance Analyzer

Affordance Analyzer: **reanalyzes** the given videos using the knowledge from the symbolic task planner to acquire the **affordance information necessary** for the **robot's effective execution**.
Specifically, it focuses on the relationship between hands and objects based on the task's nature and object names.

- Attention to Human Hands to Detect Grabbing and Releasing
- **Attention to Hand-Object Interaction to Detect the Spatiotemporal Location of Grabbing and Releasing**

The pipeline then focuses on the **grasp video clip**, analyzing the position and timing of the grasped object.

Use Detic, an off-the-shelf, open-vocabulary object detector [1], to search for object candidates within the video, as identified in the symbolic task planner.

When multiple object candidates are identified, the one closest to the hand in the video clip is deemed the grasped object.

[1] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in ECCV, 2022.

Method – Affordance Analyzer

Affordance ... lanner to acquire the afforda... Specifically ... l object names.

- Att...
- Att... nd Releasing

Th... e grasped object.

Us... wit... dates

Wh... dee... p is

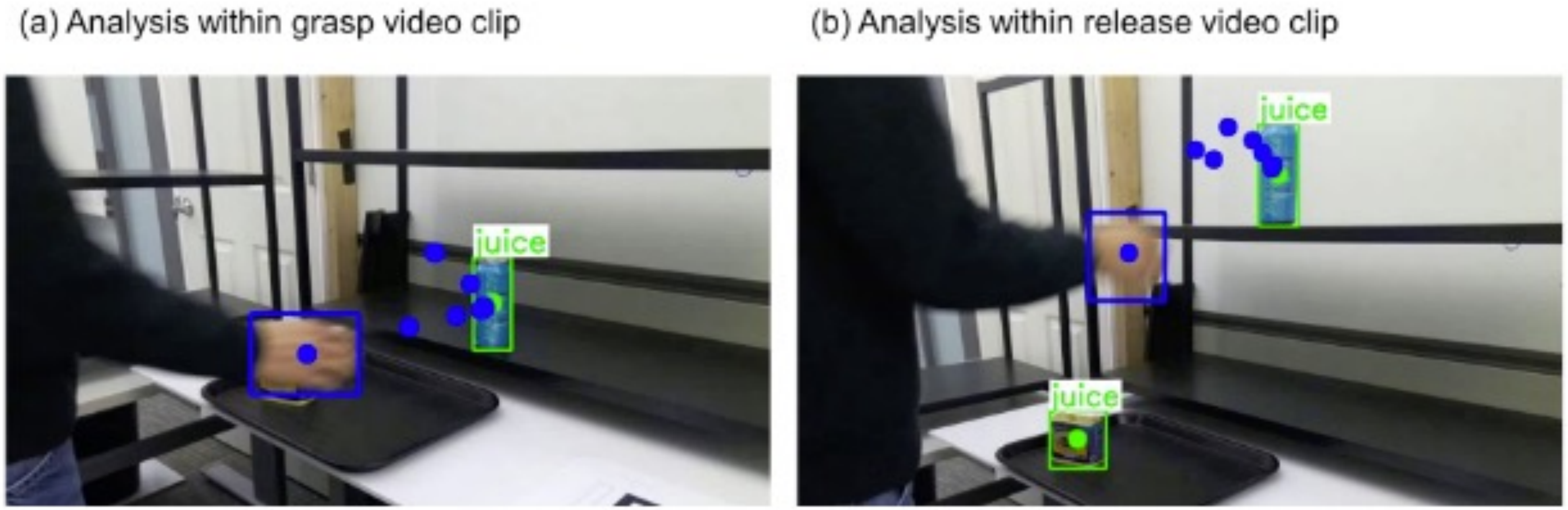


Fig. 7. Detection of the objects by focusing on the relationship between the hand and the object. The first frame and the last frame are shown for the grasp and release video clip, respectively. Green rectangles are the candidates for the object detected by the Detic model. When multiple object candidates are identified, the one closest to the hand in the video clip is deemed the grasped object. The hand positions in the video clip were illustrated as blue points. Images are spatially cropped for the visualization purpose.

[1] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in ECCV, 2022.

Method – Affordance Analyzer

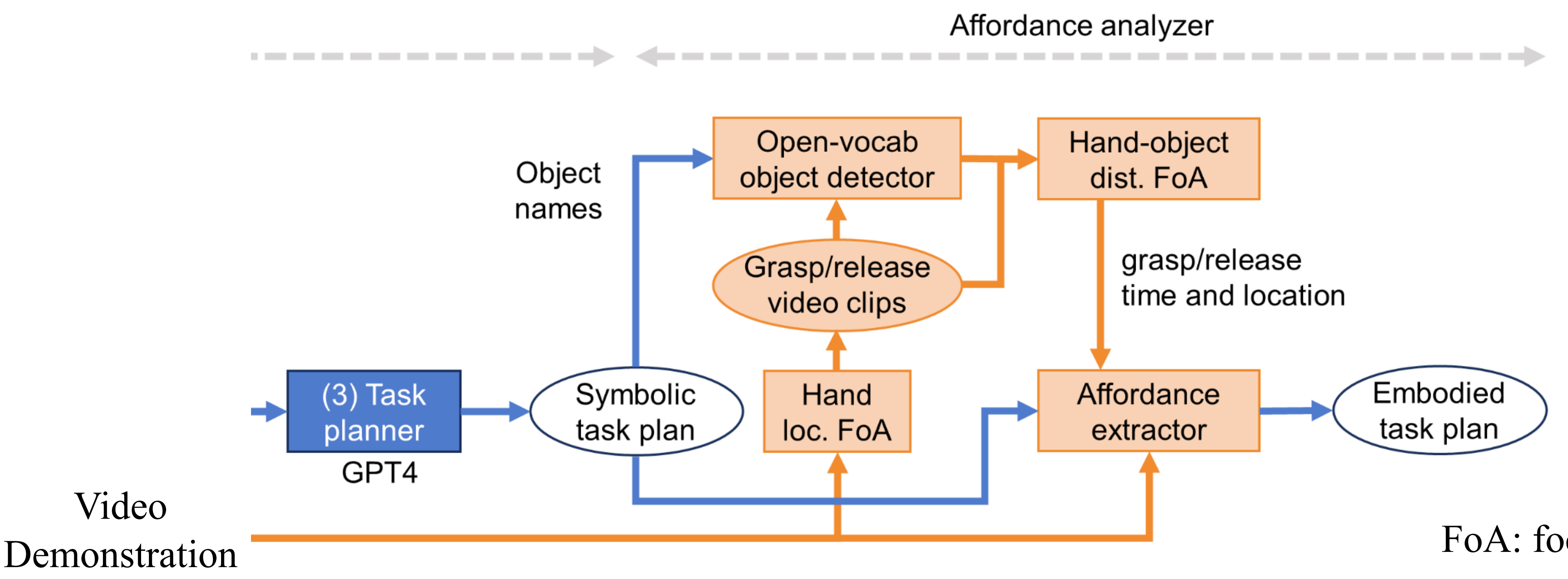
Affordance Analyzer: **reanalyzes** the given videos using the knowledge from the symbolic task planner to acquire the **affordance information necessary** for the **robot's effective execution**.
Specifically, it focuses on the relationship between hands and objects based on the task's nature and object names.

- Attention to Human Hands to Detect Grabbing and Releasing
- Attention to Hand-Object Interaction to Detect the Spatiotemporal Location of Grabbing and Releasing
- **Extracting Affordance from Aligned Videos**

1. Affordance of the Grab task: 1) Information about the approach direction towards the object to avoid collisions with the environment. 2) The grasp type also contains knowledge about how humans efficiently perform manipulations.
2. Affordance of the MoveHand task: 1) Information about waypoints during the hand's movement to avoid environmental collisions.
3. ……

Method – Affordance Analyzer

Affordance Analyzer: **reanalyzes** the given videos using the knowledge from the symbolic task planner to acquire the **affordance information necessary** for the **robot's effective execution**.
Specifically, it focuses on the relationship between hands and objects based on the task's nature and object names.



FoA: focus-of-attention

MANCHESTER 1824
The University of Manchester

Experiments

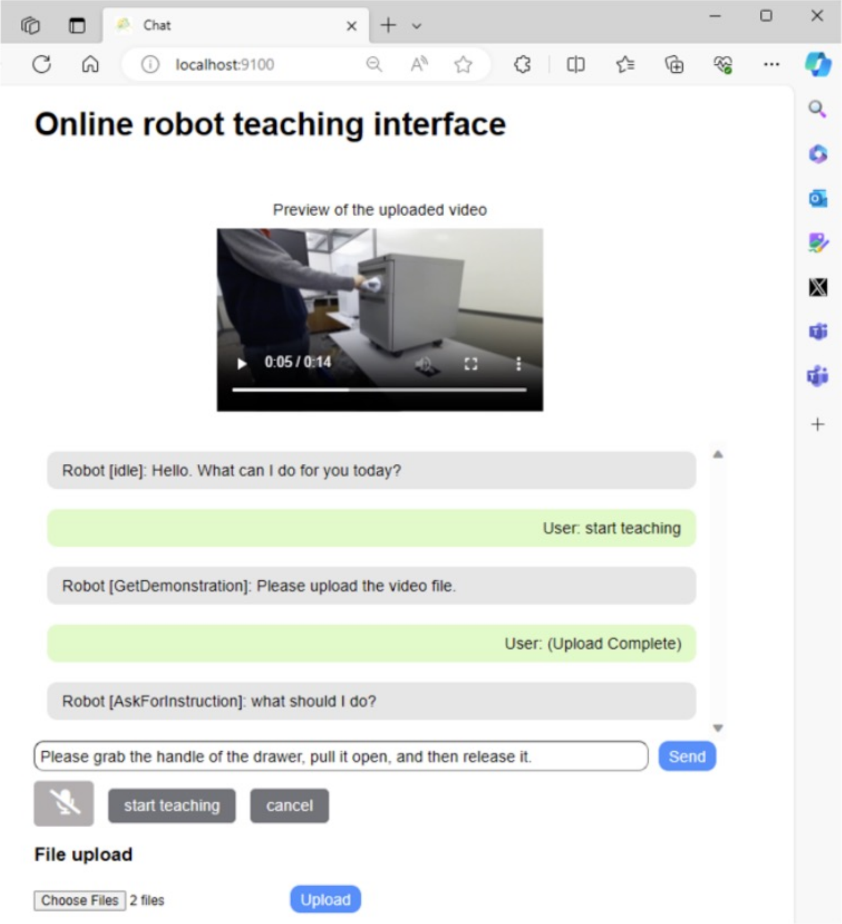


Fig. 9. A web interface to operate the proposed pipeline.

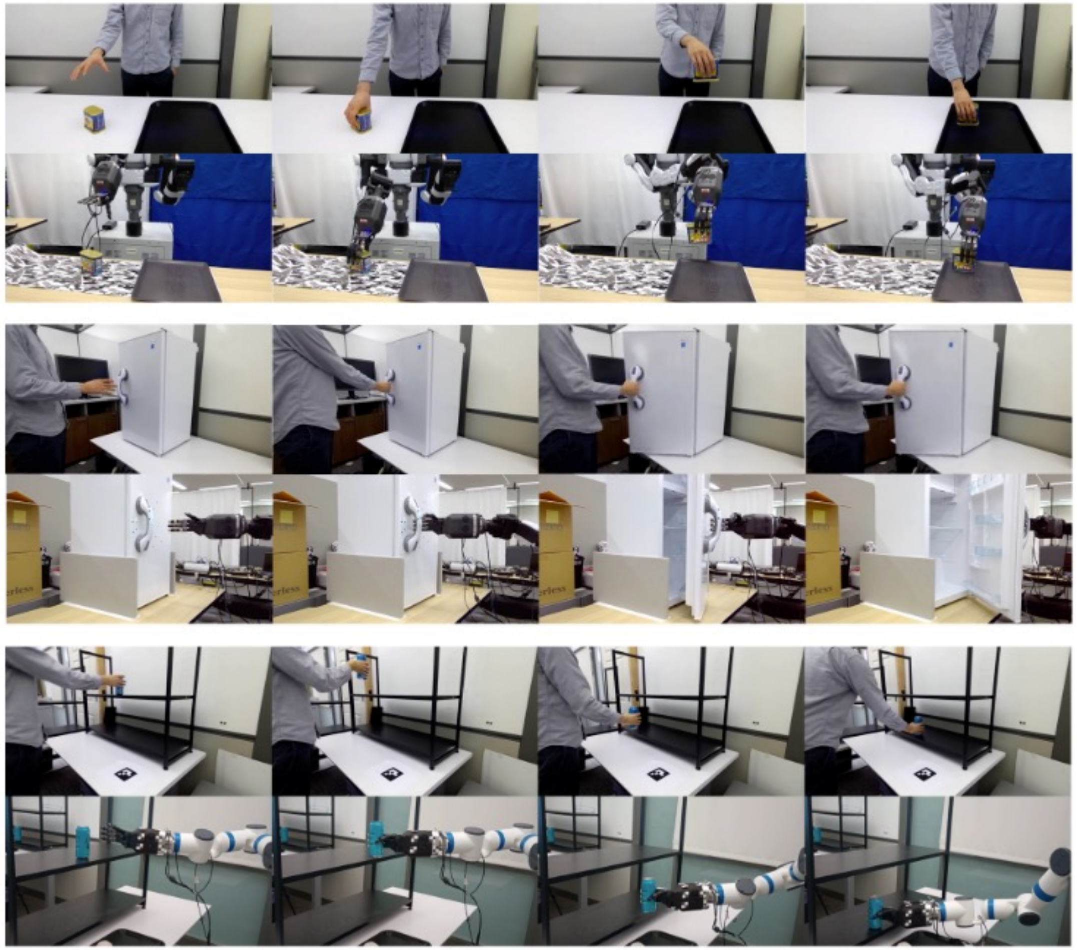


Fig. 10. Examples of the robot execution based on human demonstration data. Top pane: moving the can of spam from the desk to the tray. Bottom pane: Opening a refrigerator. Bottom pane: relocating the juice between the shelves. All the experimental results are available at

Videos: https://microsoft.github.io/GPT4Vision-Robot-Manipulation-Prompts/

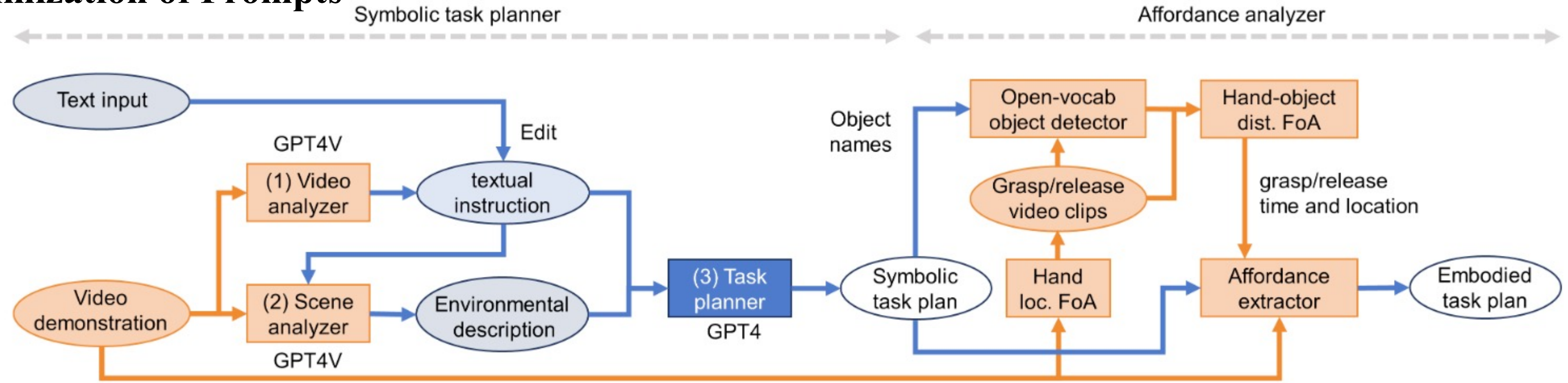
The University of Manchester

Conclusion

Introduced a novel multimodal robot task planning pipeline utilizing GPT-4V, effectively converting human actions from videos into robot-executable programs.

Limitations:

- **Extension to Long Steps:** The grounding technique was confined to grasping and releasing tasks, limiting the range of extracted affordance information.
- **Higher-order Pre- and Post-conditions:** the criteria for completing a task frequently surpass current simple object interactions. For example, a MoveOnSurface task for cleaning should not only ensure the continuous contact of the held object (like a sponge) with the surface but also the removal of dirt from that surface.
- **Optimization of Prompts**



Discussion