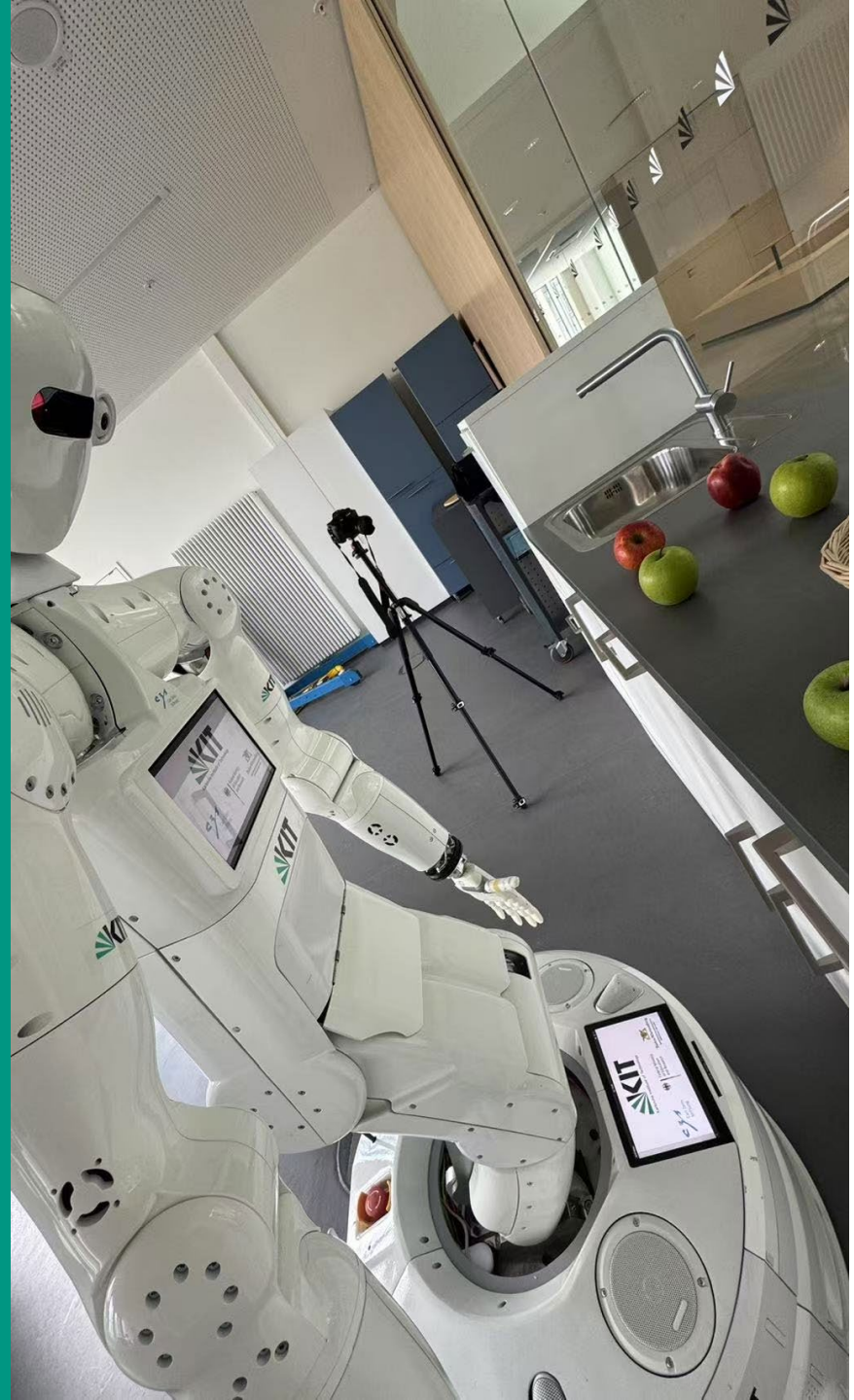
Foundation Models for Robot Task Planning



Weiheng Wang, Sheng Liu, Jiekun Chen

Supervisor: Timo Birr

KIT

H2T
High Performance Humanoid Technologies

Motivation

- **Complex and Ambiguous of Nature Language**

- **Low level semantic details can’t be fully represented by language only**

  e.g. Object positions, Spatial relationships, Visual context

- **World is multimodal**

Our Contribution

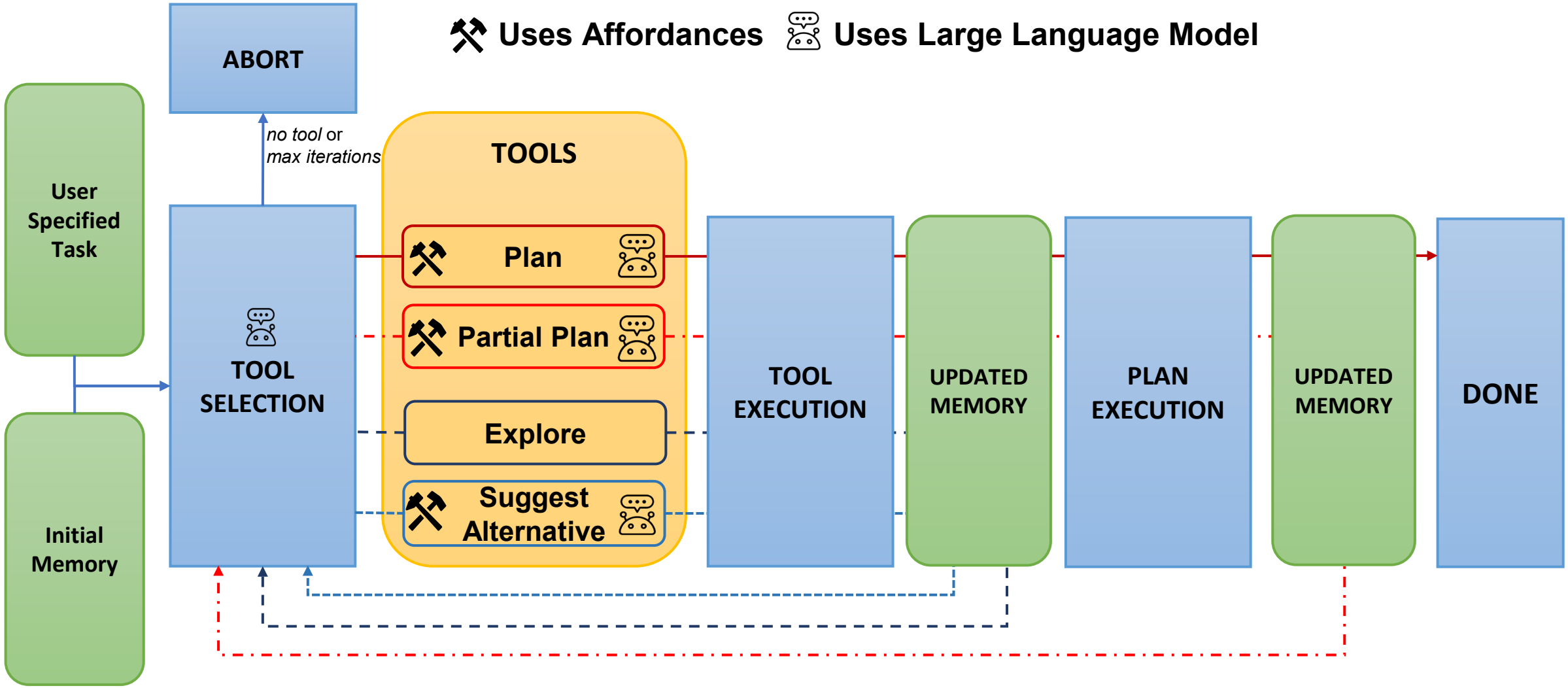
- Improve the performance of AutoGPT+P[1] with Vision-Language-Model

- Parallel Research with Image2PDDL[2](preprinted on Jan 2025)

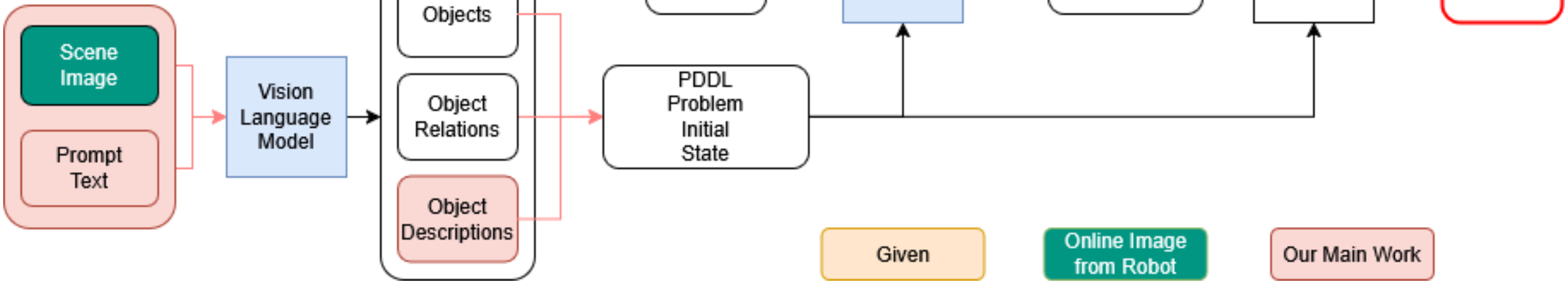
Scene

Online Image From Robot

What we want to do



What we have done



What we have done

SCENE-DESCRIPTION

OBJECTS

apple:0 #red

apple:1 #red

apple:2 #green

apple:3 #green

apple:4 #green

coffee_cup:0 #blue

bowl:0 #woven

table:0 #grey

human:0 #human

END-OBJECTS

RELATIONS

on apple:0 table:0

on apple:1 table:0

on apple:2 table:0

on apple:3 table:0

on apple:4 table:0

on coffee_cup:0 table:0

on bowl:0 table:0

at human:0 table:0

at robot:0 table:0

END-RELATIONS

LOCATIONS

table0 table:0 apple:0 apple:1 apple:2 apple:3 apple:4 coffee_cup:0 bowl:0

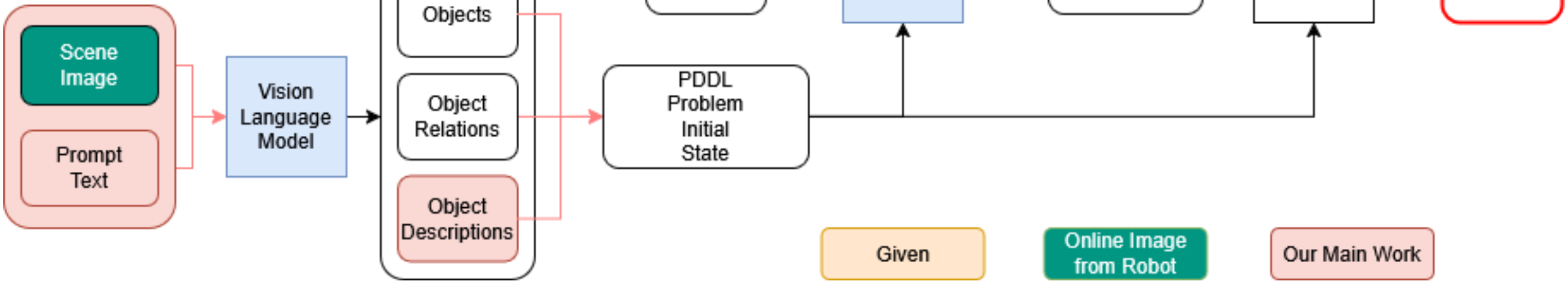
END-LOCATIONS

LOCATIONS

table0 table:0 apple:0 apple:1 apple:2 apple:3 apple:4 coffee_cup:0 bowl:0

END-LOCATIONS

What we have done



What we have done

```
(define (problem test)

   (:domain robotic_planning)

   (:objects
      apple3 - apple ; green
      robot0 - robot_profile ;
      coffee_cup0 - coffee_cup ; blue
      apple2 - apple ; green
      apple0 - apple ; red
      table0 - table ; grey
      bowl0 - bowl ; woven
      human0 - human ; human
      apple1 - apple ; red
      apple4 - apple ; green
   )
```

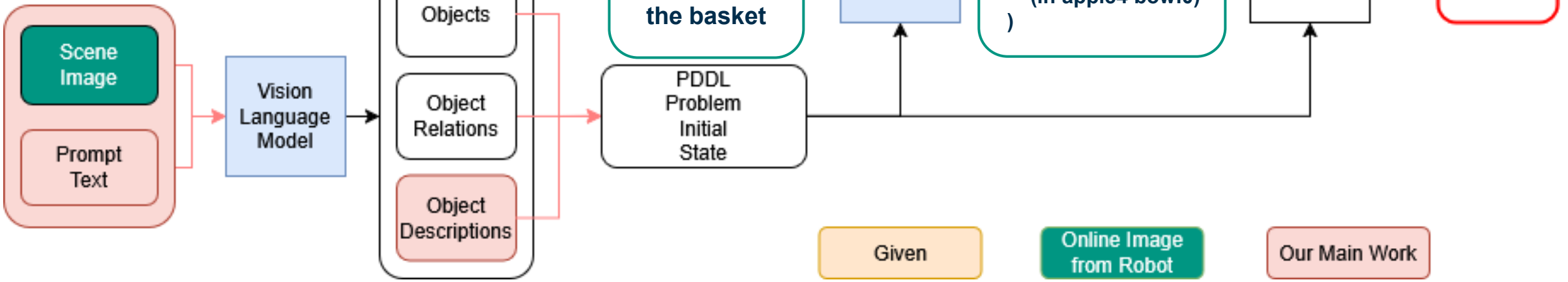
```
   (:init
      (on  apple1 table0)
      (on  apple3 table0)
      (on  apple4 table0)
      (at  human0 table0)
      (on  apple2 table0)
      (on  coffee_cup0 table0)
      (on  apple0 table0)
      (on  bowl0 table0)
      (at  robot0 table0)
      (= total-cost 0)
      (= (cost robot0) 1)
      (= (cost human0) 100)
   )

   (:metric minimize (total-cost))

)
```

What we have done

grasp robot0 apple4 table0 table0
putin robot0 apple4 bowl0 table0
grasp robot0 apple3 table0 table0
putin robot0 apple3 bowl0 table0
grasp robot0 apple2 table0 table0
putin robot0 apple2 bowl0 table0



Conclusion

→ Integrated Multimodal Approach

→ Comprehensive Scene Understanding

Future Work

→ Enhanced Multimodal Perception

→ Real-Time and Dynamic Planning

→ Interactive Feedback Loop

→ Expansion to Complex Tasks and Domains

→ Relative Position for Objects

References

[1] Birr, T., Pohl, C., Younes, A., & Asfour, T. (2024). Autogpt+ p: Affordance-based task planning with large language models. *arXiv preprint arXiv:2402.10778*.

[2] Dang, X., Kudláčková, L., & Edelkamp, S. (2025). Planning with Vision-Language Models and a Use Case in Robot-Assisted Teaching. *arXiv preprint arXiv:2501.17665*.