CasaBot
Home Assistance: The Fusion of AI and Robotics for Daily Tasks

Daniel Tan Sioa Hen: 12111127 | Ng Wooi Cheng: 12111128

汪俊扬: 12111028 | 周靖东: 12111026

毛新科:11910412 | 邓皓文: 12110510 | 季亦冰12110501

Group 4

2024.3.24

AncoraSIR.com

Proposed Project Title Summary

The project aims to develop a collaborative robot arm for home automation, leveraging natural language interaction to enhance accessibility and autonomy. Integrating robotics, AI, computer vision, and natural language processing, the initiative seeks to advance human-robot interaction, enabling real-time task execution beyond traditional automation boundaries. With a focus on aiding people with disabilities, the project drives technological innovation in AI and robotics, pushing the boundaries of machine learning to comprehend complex instructions and environmental cues. Through readings covering object recognition and manipulation techniques, including studies by Bulun, Tremblay et al., Mousavian et al., and Breyer et al., alongside resources on YOLOv8 and GPT's role in robotics, the project gains insights crucial for its development. Utilizing datasets such as LLM prompt dataset, COCO, FAT, and YCB-Video, along with simulation tools like Robosuite and Gazebo, and deep learning frameworks like PyTorch, the project focuses on object recognition using models like Faster R-CNN, SSD, and YOLOv8. Advancements in object pose detection and grasping involve deep learning methods trained on synthetic data. Deep Object Pose Estimation uses a mix of domain randomized and photorealistic data to train a neural network for accurate 6-DoF pose estimation from single RGB images. 6-DoF GraspNet employs variational autoencoders to generate and refine grasp poses using 3D point cloud data. Volumetric Grasping Network utilizes a 3D CNN to predict 6-DoF grasps from scene information represented as a Truncated Signed Distance Function (TSDF), providing grasp quality, orientation, and width predictions at the voxel level. Integration of user APIs with ROS enables seamless interaction, facilitating tasks such as fine-tuning pre-trained models and executing actions based on commands. Natural language interaction is facilitated through text-to-speech and speech-to-text functionalities, with temperature parameter adjustment influencing output randomness during text generation. Evaluation of results encompasses simulator experiments like Cleanup and Sort Tasks, categorized by general categories, object attributes, and functions, alongside interaction experiments with volunteers to assess performance in terms of response time, task accuracy, command understanding, completion rate, and user satisfaction, ultimately pioneering the future of home automation and human-robot interaction.

What is the problem that you will be investigating?

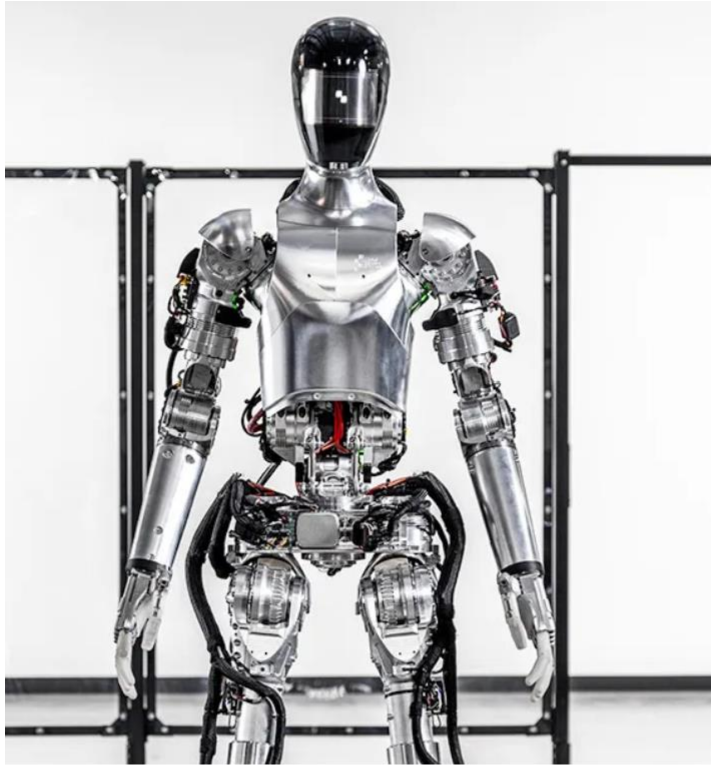
Enhancing Home Automation with AI

- Project Goal: To create a collaborative robot arm that enhances home automation by understanding and executing tasks through natural language interaction.
- Interdisciplinary Challenge: Merging robotics, artificial intelligence, computer vision, and natural language processing to understand verbal commands and visual information.
- Human-Robot Interaction: Advancing seamless interaction, making technology more accessible and intuitive.
- Home Assistance Revolution: Introducing dynamic, real-time task execution beyond traditional predefined automation, supporting a wide range of household activities.

What is the problem that you will be investigating?

Impact and Innovation

- Accessibility and Independence: Providing invaluable assistance to individuals with disabilities, offering greater autonomy within their homes.
- Technological Breakthroughs: Pushing the boundaries of AI and robotics to understand complex instructions and environmental cues, leading to innovative machine learning advancements.
- Future of Home Automation: Pioneering the next generation of home automation, where robots act as versatile assistants, transforming everyday life and interactions with technology.



AncoraSIR.com

What reading will you examine?

Object Recognition:

- Bulun, S. (2023). Object Recognition and Tracking of Bolts: A Comparative Analysis of CNN Models and Computer Vision Techniques : A Comparison of CNN Models and Tracking Algorithms (Dissertation).
- Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-337061
- https://learnopencv.com/ultralytics-yolov8/

Object Pose Detection and Grasping:

- Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., & Birchfield, S. (2018). Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790.
- Mousavian, A., Eppner, C., & Fox, D. (2019). 6-dof graspnet: Variational grasp generation for object manipulation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2901-2910).
- Breyer, M., Chung, J. J., Ott, L., Siegwart, R., & Nieto, J. (2021, October). Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In Conference on Robot Learning (pp. 1602-1611). PMLR.

GPT-part:

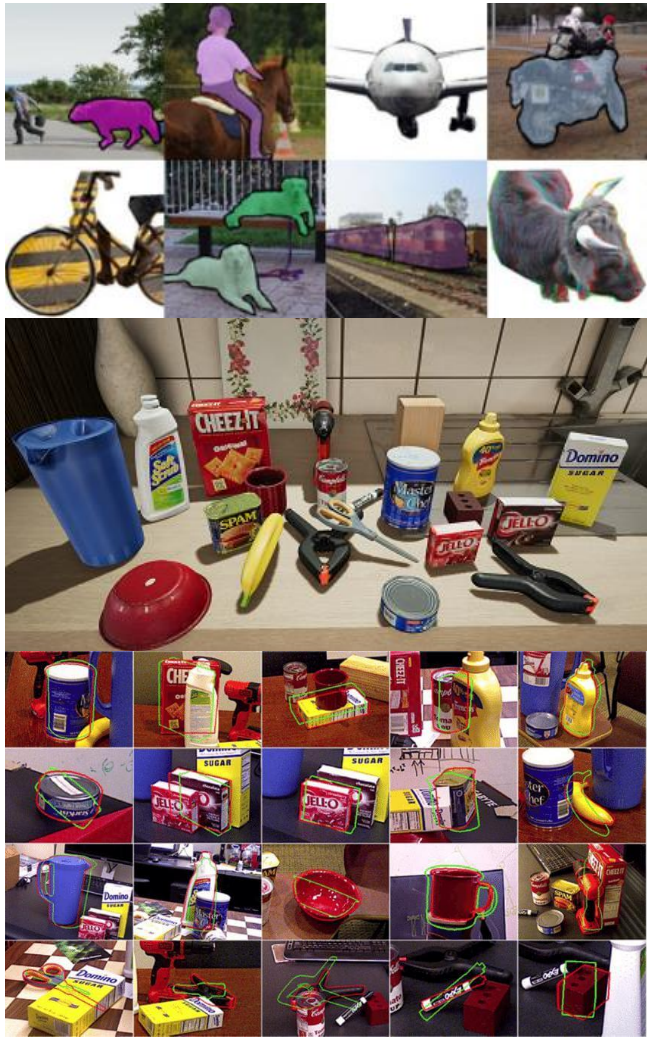
- https://www.microsoft.com/enus/research/uploads/prod/2023/02/ChatGPT___Robotics.pdf
- https://tidybot.cs.princeton.edu/
-

AncoraSIR.com

SUSTech
Southern University of Science and Technology

What data will you use?

- LLM prompt dataset from TidyBot
- Common Objects in Context (COCO)
- Falling Things(FAT),from NVIDIA
- YCB-Video dataset



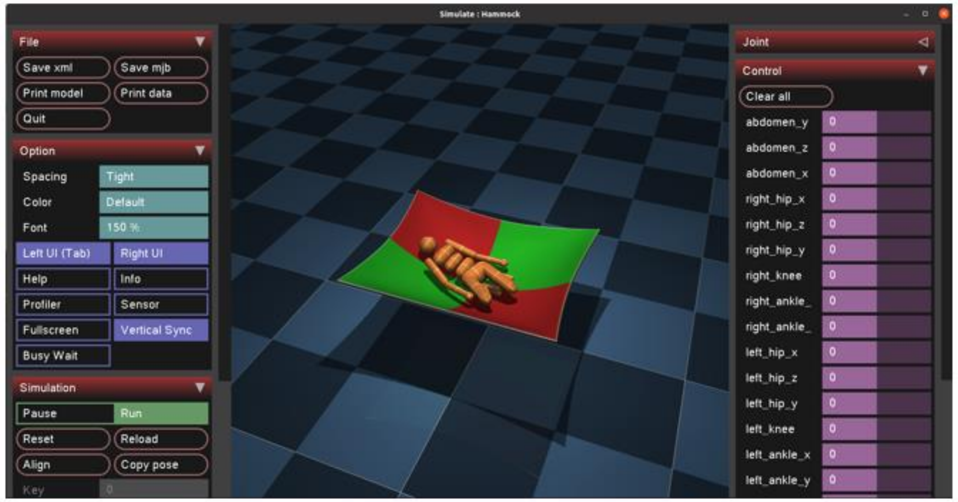
AncoraSIR.com

What method or algorithm are you proposing?

If there are existing implementations, will you use them, and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.

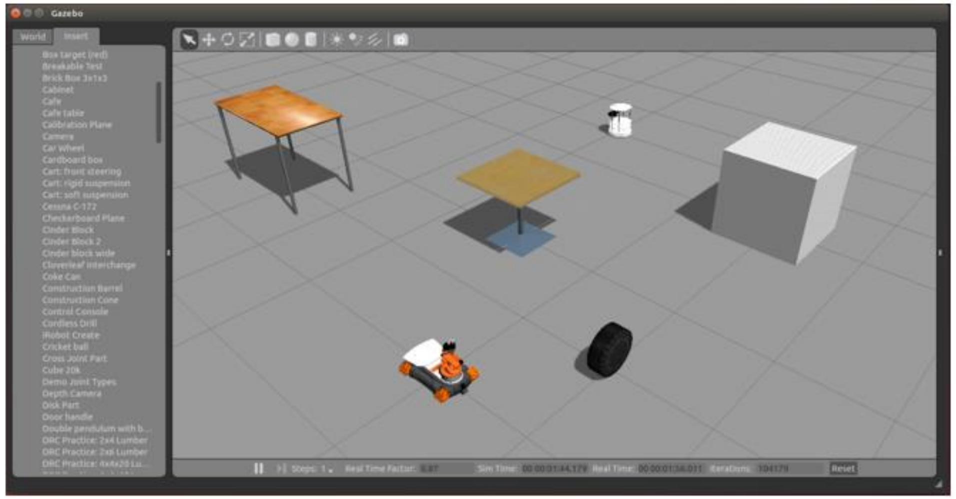
Select Simulated Environments

Robosuite



Based on MuJoCo, More userfriendly, Intuitive user interface,Higher learning curve

Gazebo



Based on ROS,Rougher models and interface,but,We have basic usage experience

The selection of the robotic arm model will be determined in the subsequent stages, with the UR5 robotic arm tentatively chosen for now.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

What method or algorithm are you proposing?

If there are existing implementations, will you use them, and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.

Object Recognition.

PyTorch is a popular deep learning framework that is frequently used for implementing object detection algorithms.

PyTorch offers a wide range of pre-implemented deep learning models for object detection, such as Faster R-CNN, SSD, and YOLO (YOLO v8).

What method or algorithm are you proposing?

If there are existing implementations, will you use them, and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.

Object Recognition.

	Pros	Cons
Faster R-CNN (Regions with Convolutional Neural Networks)	• High accuracy • Flexibility in model design	• Slower inference speed • Complex to implement and train
SSD (Single Shot MultiBox Detector)	• Faster inference speed • Easier to implement (relatively simple architecture compare to Faster R-CNN)	• Difficulty in handling aspect ratios
YOLO v8	• High speed & accuracy (suitable for real-time application) • Simple and easy to implement, making it accessible to a wide range of users	• Requires large datasets for training

What method or algorithm are you proposing?

If there are existing implementations, will you use them, and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.

Object Recognition.

Model	mAP50 [%]	Average inference time [ms]
YOLOv8n	55.6	12.72
YOLOv8n reduced	66.0	10.97
SSD	52.1	52.63
SSDLite	41.2	20.03
Faster R-CNN	40.6	350.25

Bulun, S. (2023). Object Recognition and Tracking of Bolts: A Comparative Analysis of CNN Models and Computer Vision Techniques : A Comparison of CNN Models and Tracking Algorithms (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-337061

The evaluation from the models shows that the YOLOv8, specifically the YOLOv8n model without the P5 layer in the head, performs the best in terms of average inference speeds of 10.97 ms and highest mAP50 score of 66.0%

mAP50: Mean average precision calculated at an intersection over union (IoU) threshold of 0.50. It's a measure of the model's accuracy considering only the "easy" detections.

AncoraSIR.com

What method or algorithm are you proposing?

Object Pose Detection and Grasping:

- Deep Object Pose Estimation

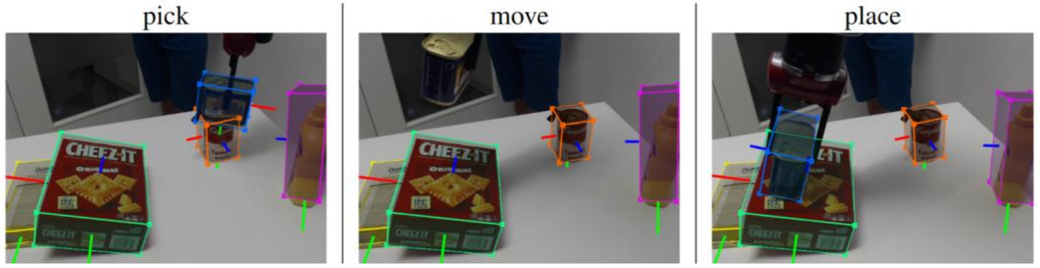
Based on synthetic data, using a combination of domain randomized and photorealistic synthetic data to train a deep neural network for 6-DoF pose estimation of known objects from a single RGB image.

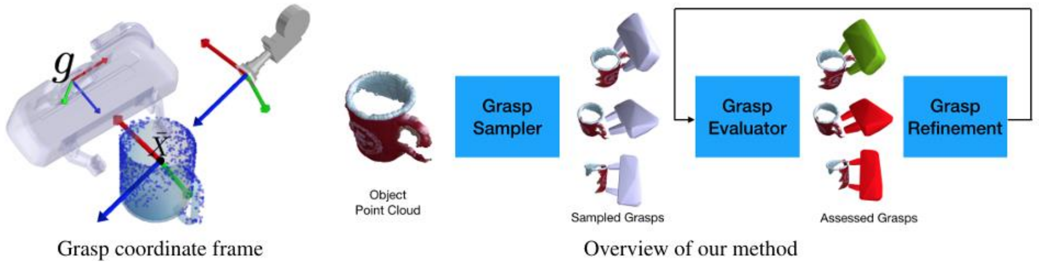
- 6-DoF GraspNet

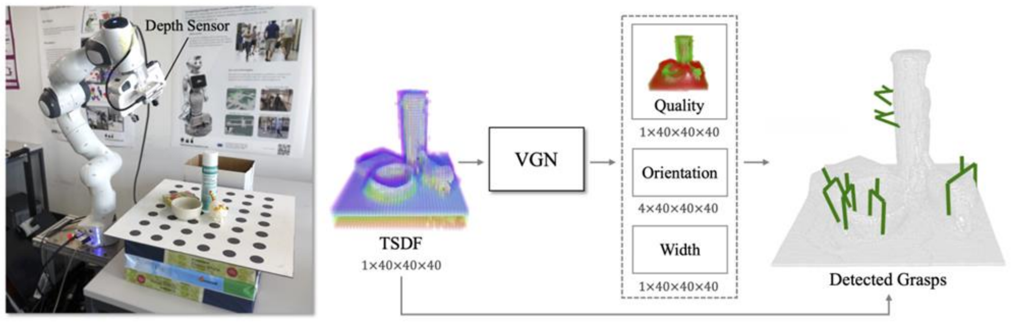
Employing a method based on variational autoencoders to generate grasp poses. This approach sampling a set of grasp poses using variational autoencoders and utilizes a grasp evaluation model to assess and refine these sampled grasps, based on 3D point clouds observed by depth cameras.

- Volumetric Grasping Network

Designing and training a 3D CNN to predicts 6 DOF grasps from 3D scene information. The network takes a Truncated Signed Distance Function (TSDF) representation of the scene as input and outputs a volume of the same spatial resolution. Each cell within this volume contains the predicted quality, orientation, and width of a grasp executed at the center of the voxel.







AncoraSIR.com

What method or algorithm are you proposing?

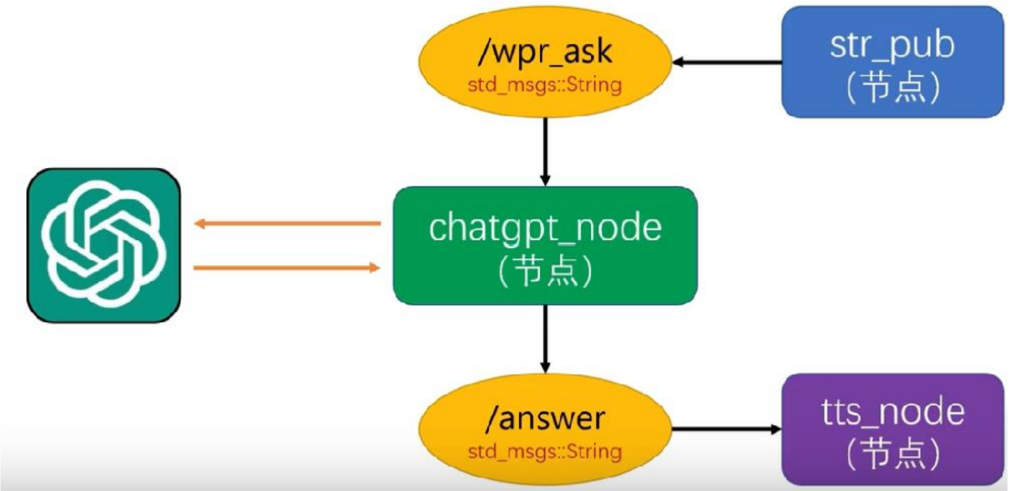
Object Pose Detection and Grasping：

	Pros	Cons
Deep Object Pose Estimation (DOPE)	• Simple input (RGB image) • Accurate estimation of the 6D pose of objects • Various scenes	• Performance might degrade in cluttered or complex scenes • May not directly provide grasp planning
6-DoF GraspNet	• Providing precise 6-DoF grasp poses • Utilizes synthetic data for training • Can handle cluttered scenes and unknown objects • Trained in simulation and can be directly applied in the real world	• No direct object pose estimation • Noise sensitive • Data dependent
Volumetric Grasping Network (VGN)	• real-time 6 DOF grasp detection • high success rates in cluttered scenes without explicit collision checking	• Relatively complex architecture and training process (TSDF) • Data dependent

AncoraSIR.com

What method or algorithm are you proposing?

- Calling user's API to integrate with ROS.

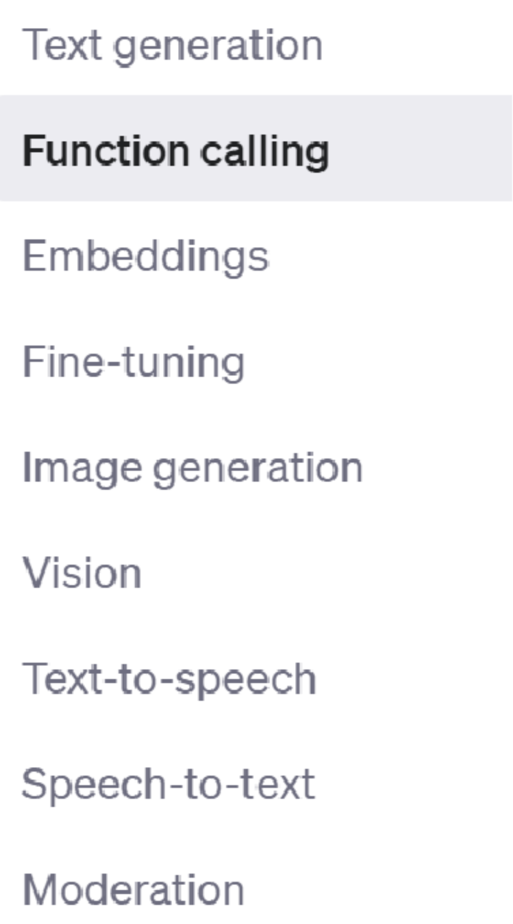



```python
from openai import OpenAI

client = OpenAI()

stream = client.chat.completions.create(
    model="gpt-4",
    messages=[{"role": "user", "content": "Say this is a test"}],
    stream=True,
)
for chunk in stream:
    if chunk.choices[0].delta.content is not None:
        print(chunk.choices[0].delta.content, end="")
```

https://github.com/play-with-chatgpt/wpr_chatgpt

AncoraSIR.com

What method or algorithm are you proposing?



1.Fine-tuning: Customizes a pre-trained model to better suit specific tasks by training on a specialized dataset. CasaBot could use fine-tuning to improve its understanding of household-specific language and tasks and reply us in a specific format.

2.Function Calling: Executes specific tasks or actions in response to commands. For CasaBot, this allows it to interpret commands and call the appropriate routines for actions like grasping objects.

3. Vision: Refers to the machine's ability to understand and interpret visual data. For CasaBot, computer vision is essential for identifying objects and navigating the home environment.

4. Text-to-Speech &Speech-to-Text: Interactions.

5.Temperature:A parameter used during text generation that influences the randomness of the output.

What method or algorithm are you proposing?

Simple Test Without API

i am the owner of this room. command: i do not like this chair

Given your command that you do not like the chair in the image, as the brain of the robotic arm, here's how I might respond using the provided functions:

1. `frame = read_camera_frame()`: Capture the current view of the room.
2. `objects = detect_objects(frame)`: Identify all objects, including the chair.
3. Locate the chair among the detected objects.
4. Calculate a path to move the chair without colliding with other objects.
5. `move_arm(x_pos, y_pos, z_pos)`: Position the arm above the chair.
6. `set_gripper_width(width)`: Adjust the gripper to match the chair's dimensions.
7. `grasp(force)`: Firmly grip the chair.
8. `move_arm(new_x_pos, new_y_pos, z_pos)`: Carry the chair to a new location or to where it can be removed from the room.
9. `release_grip()`: Release the chair gently at the new location.
10. `move_arm(home_x, home_y, home_z)`: Return the arm to a neutral position.

Please note, the actual x, y, z coordinates and width would be specific values calculated based on the room's layout and the chair's size and position. This script assumes the robotic arm has the necessary strength and reach to move the chair, and that the environment is safe for such an operation.

What method or algorithm are you proposing?

Simple Test Without API

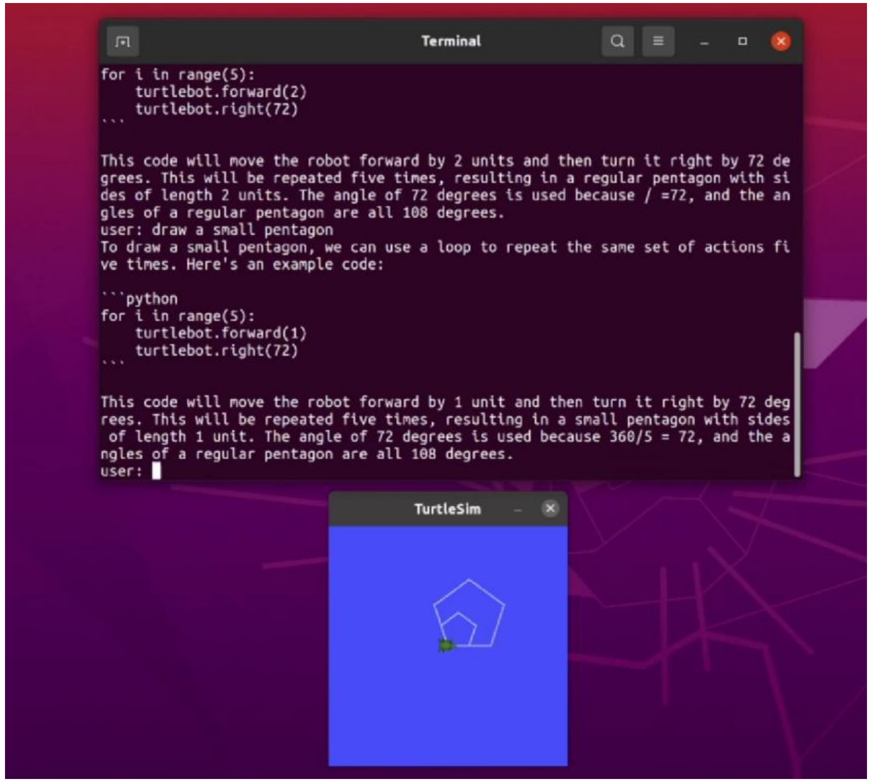
its time for dinner. but where is my friend

```python
frame = read_camera_frame()
objects = detect_objects(frame)
# Assuming 'prepare_table' is a function to set the table for dinner
prepare_table = get_object_position(objects['table_setting'])
move_arm(prepare_table.x, prepare_table.y, prepare_table.z)
set_gripper_width(prepare_table.width)
grasp(prepare_table.force)
# Place an additional setting for the friend
new_setting_position = calculate_new_position_for_setting()
move_arm(new_setting_position.x, new_setting_position.y, new_setting_position.z)
release_grip()
# Return to a waiting position, possibly programmed as a 'home' or 'idle' state for t
move_arm(home_position.x, home_position.y, home_position.z)
```

AncoraSIR.com

What method or algorithm are you proposing?

Related Works



https://github.com/chinaheyu/chatgpt_ros

TidyBot

Model	Commonsense		Summarization	
	Seen (%)	Unseen (%)	Seen (%)	Unseen (%)
text-davinci-003	45.0	45.6	**91.8**	**91.2**
text-davinci-002	41.8	37.5	84.1	75.7
code-davinci-002	41.4	39.4	88.6	83.2
PaLM 540B	**45.5**	**49.6**	84.6	75.7

https://rdcu.be/dChXj

AncoraSIR.com

How will you evaluate your results?

Experiment in Simulator

Cleanup Task

Two different scenarios, each with its own set of 10 objects, 2-3 receptacles.

Sort Task:

- based on general categories (e.g.,put clothes here and toys there)
- based on object attributes (e.g., put plastic items here and metal items there)
- based on function (e.g., put winter clothes here and summer clothes there)

Method	Category (%)	Attribute (%)	Function (%)
Examples only	80.1	72.7	75.7
WordNet taxonomy	69.1	59.8	61.4
RoBERTa embeddings	78.6	75.5	71.8
CLIP embeddings	84.6	79.8	85.5
Summarization (ours)	**91.0**	**85.6**	**93.9**

AncoraSIR.com

SUSTech
Southern University of Science and Technology

How will you evaluate your results?

Interaction Experiment

Invite volunteers to test our CasaBot and evaluate it from following aspects:

- **Response Time**: How quickly does CasaBot acknowledge and begin executing a command?
- **Task Accuracy**: How accurately does CasaBot perform the tasks given in the commands?
- **Command Understanding**: Does CasaBot correctly understand the details of the command?
- **Completion Rate**: What percentage of tasks is CasaBot able to complete successfully?
- **User Satisfaction**: How satisfied are users with the interactions?

SUSTech Design + Learning Lab PowerPoint Template

Name

Affiliation

Supervisor

Contact

AncoraSIR.com