Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface

Future Directions

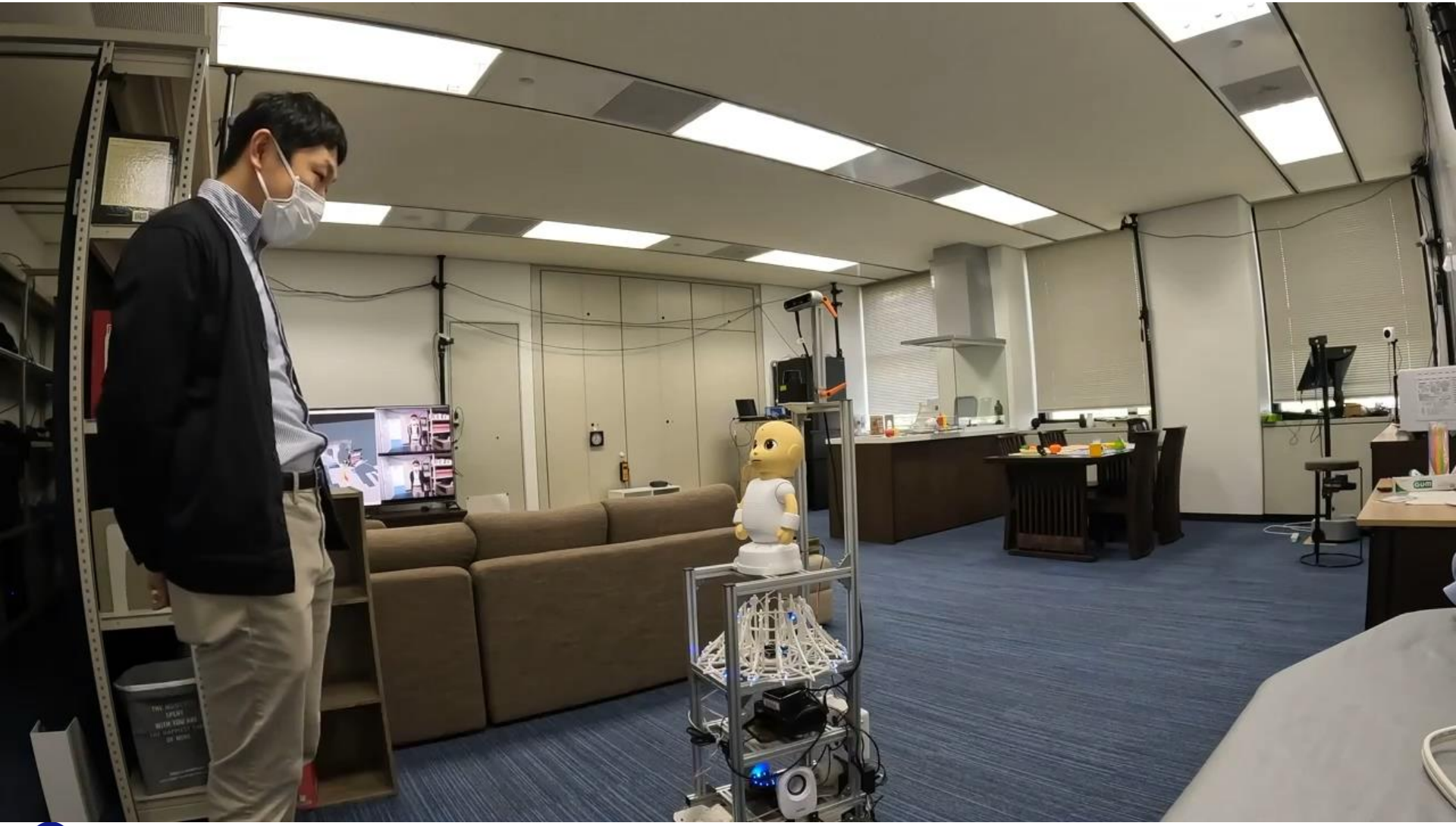
Yutaka Nakamura, Shuhei Kurita, Koichiro Yoshino
Guardian Robot Project, RIKEN
Center for Advanced Intelligence Project, RIKEN

知識獲得・対話研究チーム
Knowledge Acquisition & Dialogue Research Team

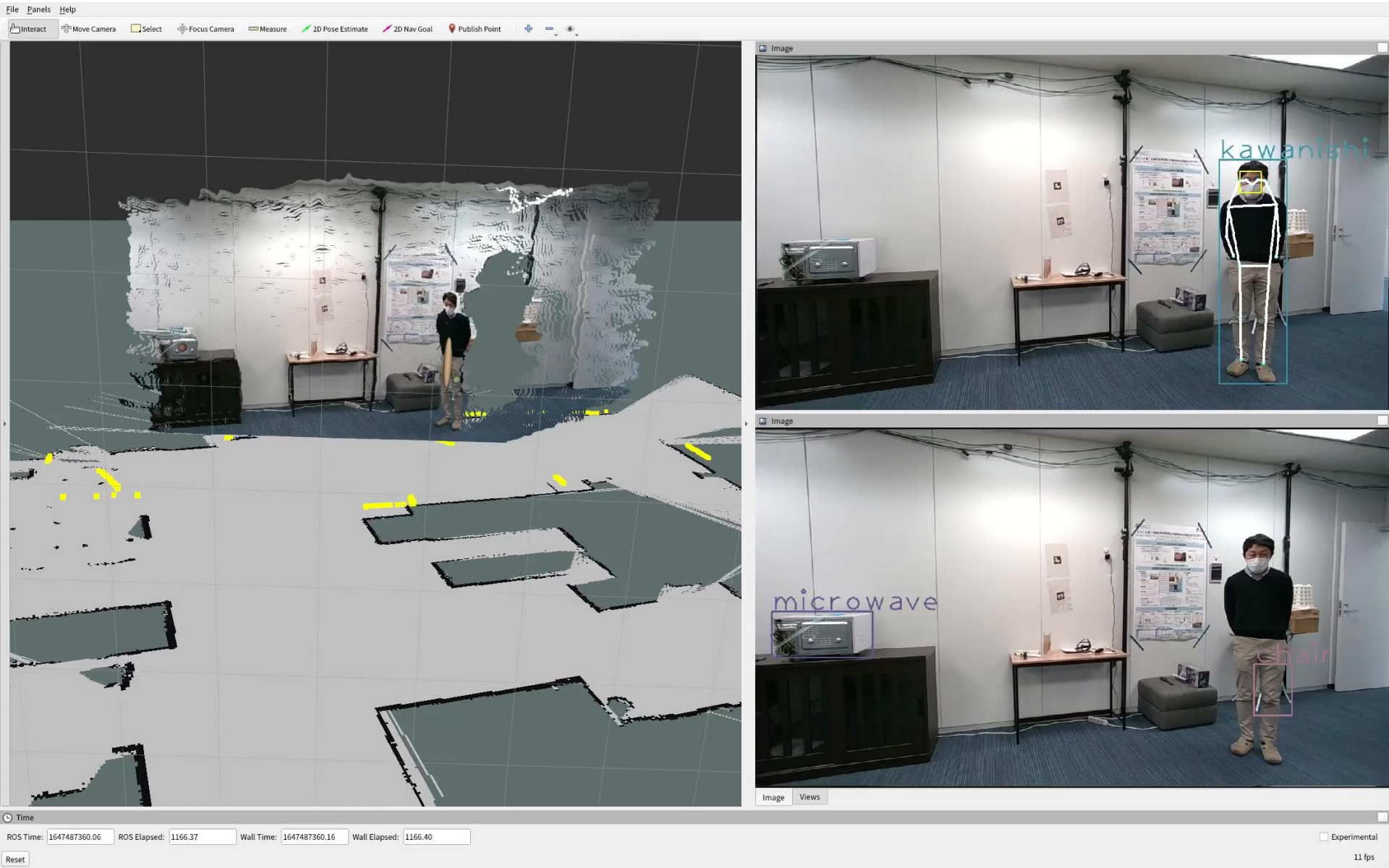
ガーディアンロボットプロジェクト
Guardian Robot Project

革新知能統合研究センター
Center for Advanced Intelligence Project

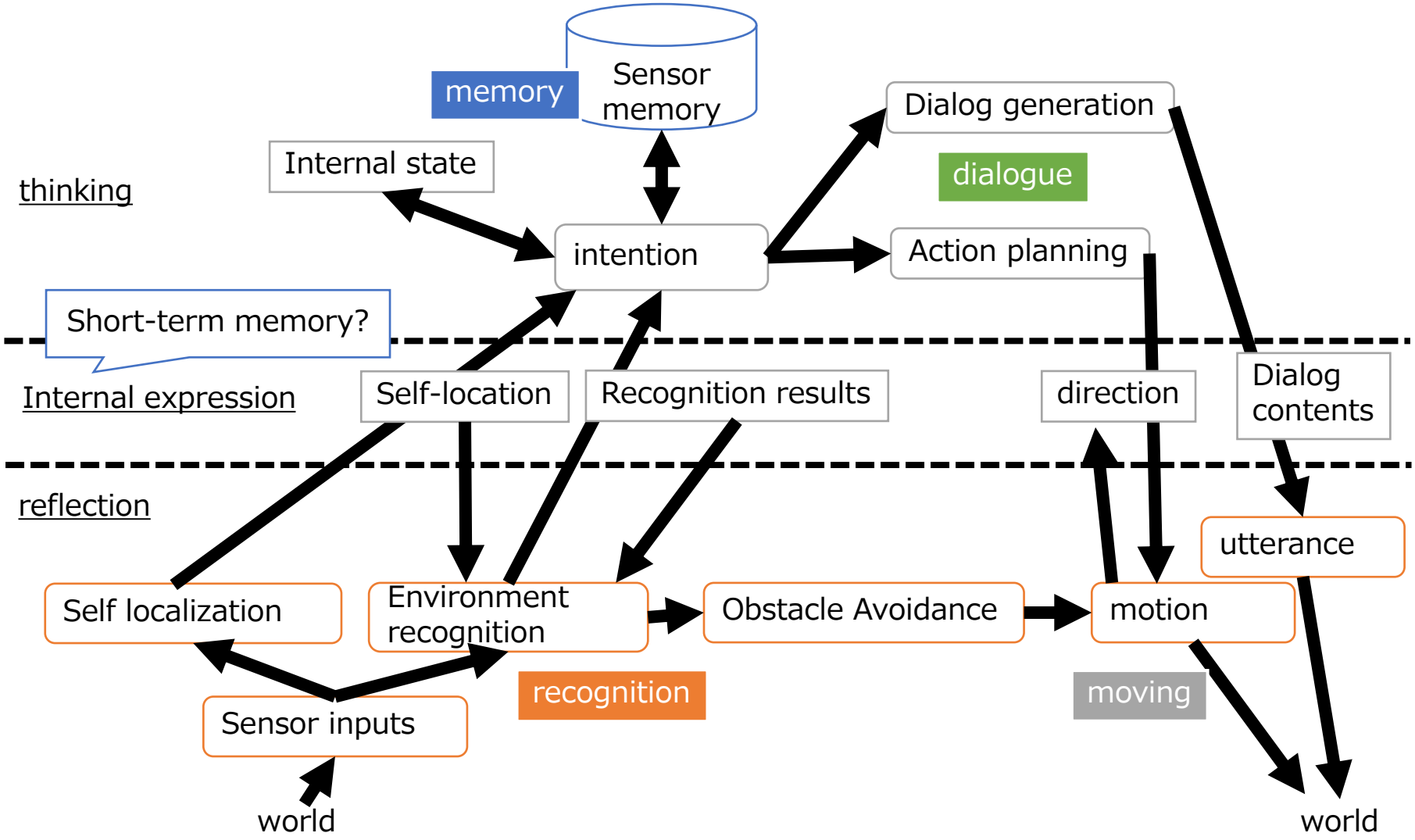
Understanding the surrounding situation



The first-person view

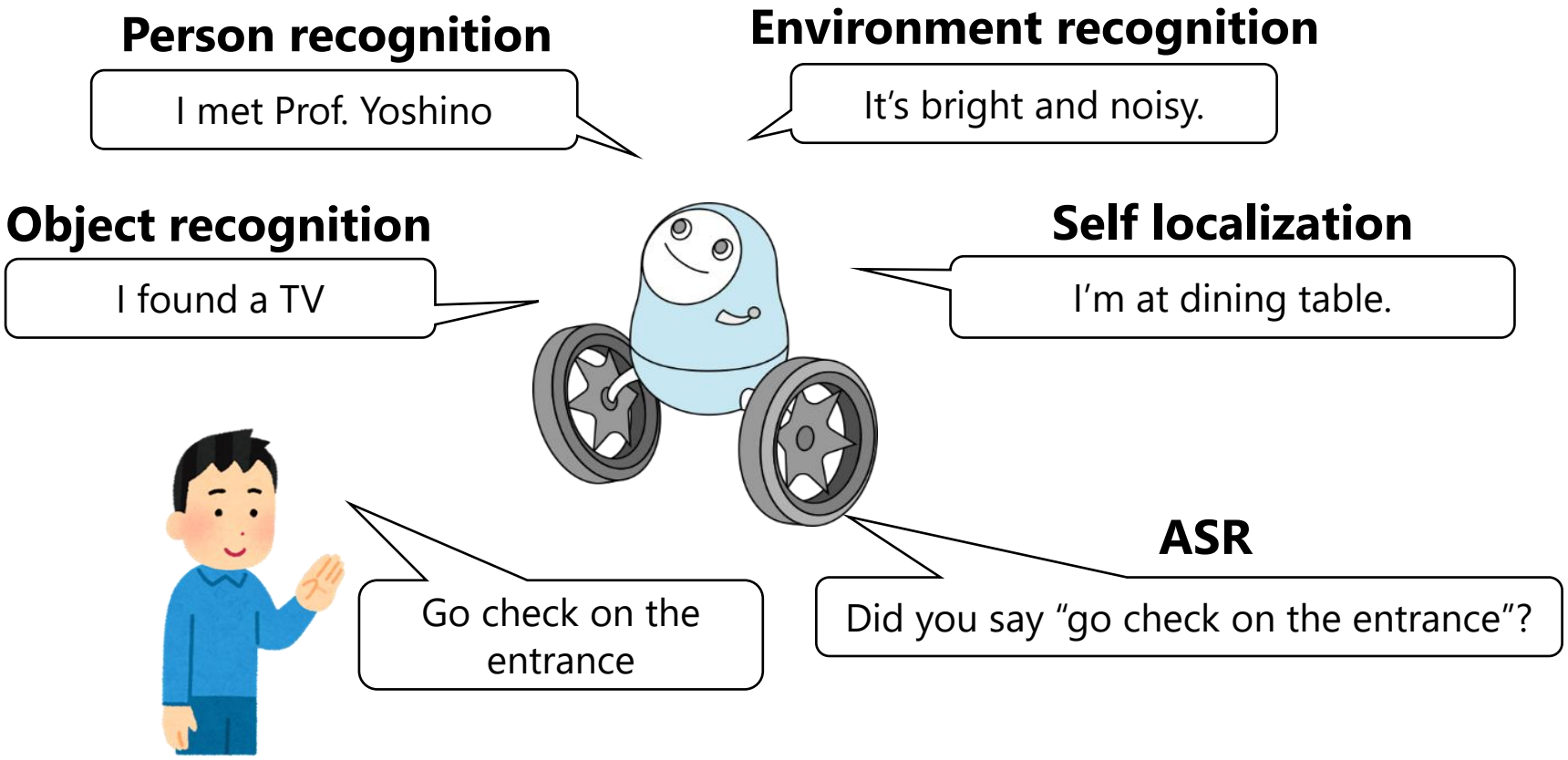


System integration empowered by Robot Operating System (ROS2)



Conversational ability of robot

◆"What to say" is important for patrolling robot



Robot evaluation

- **Operability:** Did you feel that the robot was doing its job of patrolling properly? (**1-5, 5 is the best**)
- **Appropriateness:** Did you feel that the robot was able to appropriately explain what it saw, heard, and felt? (**1-5, 5 is the best**)
- **Individuality:** Did you feel that the robot acted based on its own will and judgment? (**1-5, 5 is the best**)
- **Conversational ability:** Did you feel that the robot was communicating appropriately with people? (**1-5, 5 is the best**)
- **Sufficiency:** Did you feel that the explanation of the robot was sufficient and necessary? (**1-7, 4 is the best**)

Subjective evaluation

◆Patrolling + conversation

- 60 participants (36 males and 24 females)
- "Recog." system described any recognition results except ASR
- "ASR" system repeated what the user said
- "Conv." system has conversational reaction to user speech

	Recog.+Conv.	Recog.+ASR +Conv.	Recog. (w/o Conv.)	Recog.+ASR (w/o Conv.)	Silence
Operability	**3.94**	3.85	3.79	3.67	2.52
Appropriateness	3.81	**3.85**	3.79	3.73	1.42
Individuality	3.63	**3.94**	3.09	3.27	2.30
Conversation	**3.88**	3.58	2.03	3.33	1.82
Sufficiency (1-7)	4.75 (+0.75)	5.37 (+1.37)	**4.48 (+0.48)**	5.00 (+1.00)	1.39 (-2.61)

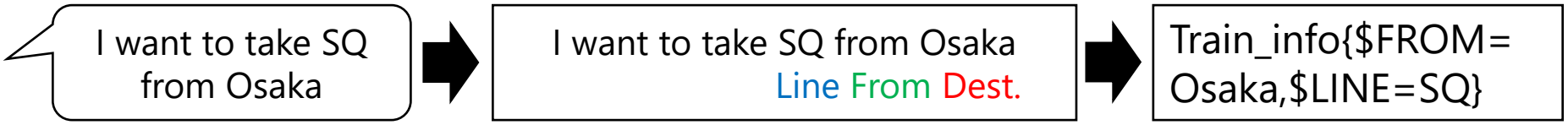
Required abilities of robots

- **Understanding well "what the user said", and "what situation the user is in."**
  - "Grounding" between observation and knowledge (language)
- **Having individuality**
  - Robots should determine necessary actions to maximize the expected satisfaction of users
    - e.g., persuasion, negotiation
- **Responding users appropriately, sufficiently, and necessarily**
  - Explanation by language and other modalities
    - Logical argumentation, emotional expressions

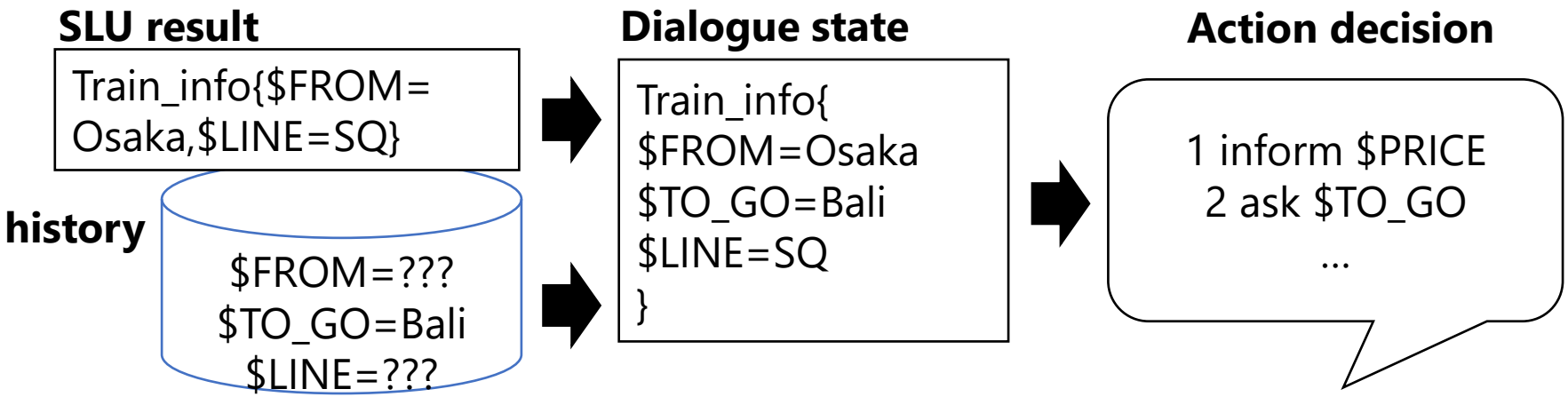
Language interpretation

◆**How does the robot interpret language?**

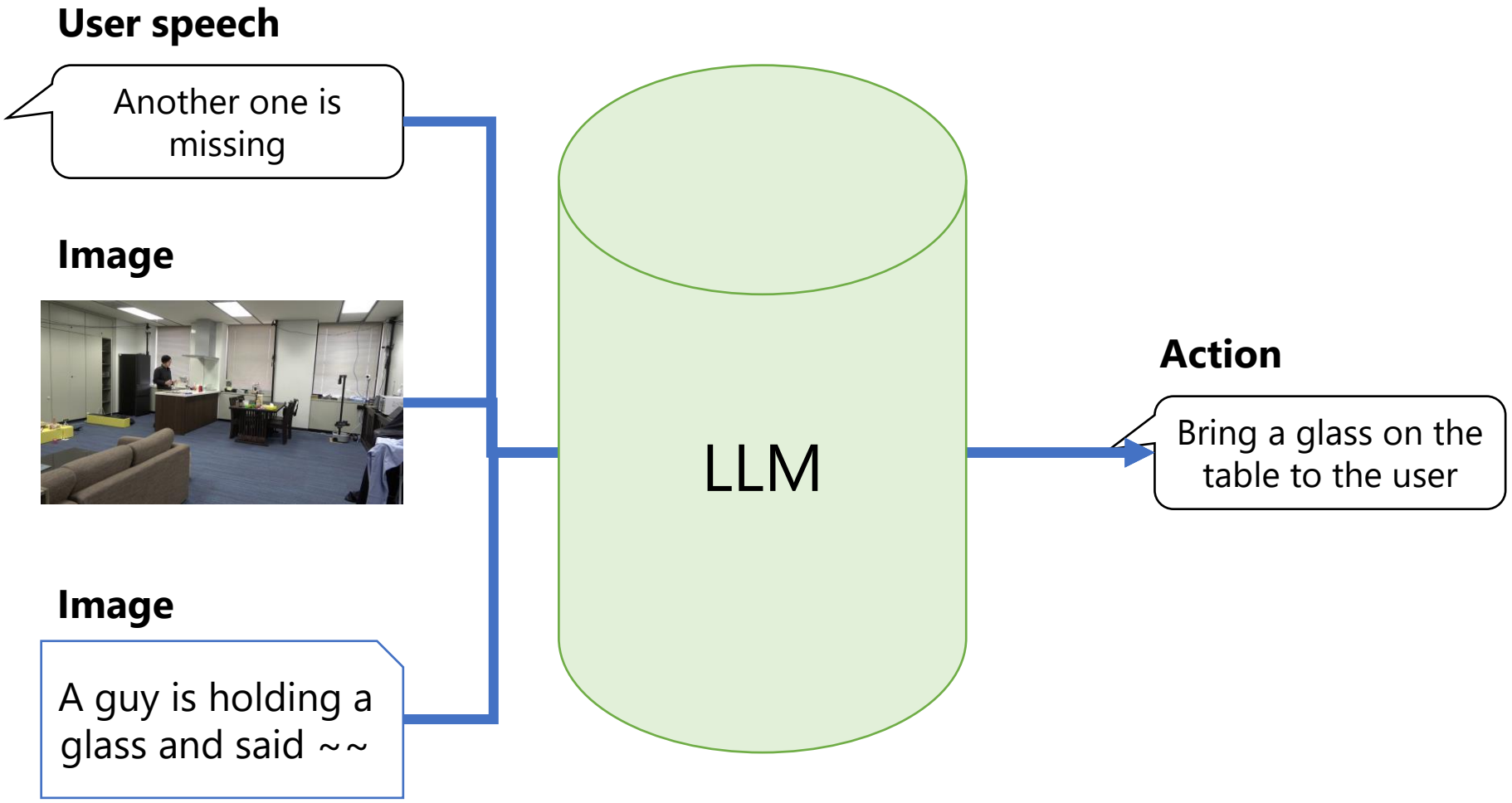
◆Spoken language understanding



◆Dialogue state tracking and action decision



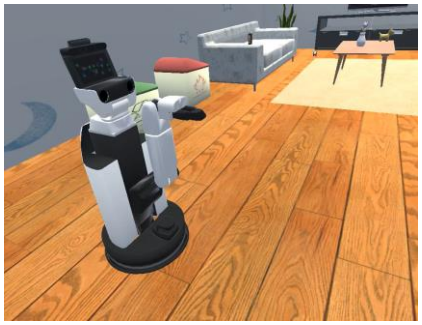
LLMs augment robot abilities



LLMs in RoboCup at home

Problem of planning

Actuation



Language (direction, description)

Taking the rabbit ornament on the kitchen table

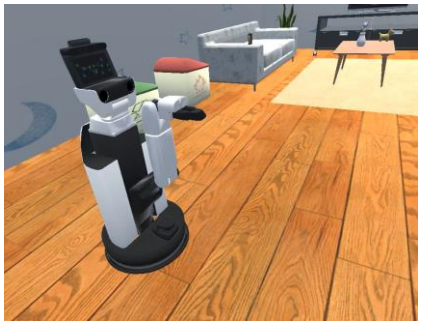
Yoshino et al., "Caption Generation of Robot Behaviors based on Unsupervised Learning of Action Segments." Proc. IWSDS, 2020.

◆Bridging actuation and language sequences

- Can the robot disambiguate the language sequence to convert them into an actuation sequence?
- Is the generated actuation sequence operative?
- Can the system consider the (physical) interaction with real-world objects in the planning?

Idea: using motion primitives

Actuation



Language (direction, description)

Taking the rabbit ornament on the kitchen table

Assuming a map

Motion primitives (a sequence basic action units)

MOVE from START to KITCHEN → PICK RABBIT ORNAMENT
→ HOLD RABBIT ORNAMENT → MOVE from KITCHEN to START

Convert a language to motion primitives

Robots generating actions by LLMs

Saycan by Google

TidyBot by Princeton & Stanford

Place clothes into the laundry basket

- **They prepare a set of motion primitives and define the planning problem as a sequence conversion from language commands to motion primitive sequences**
  - LLMs are good at sequence conversion

Generating action sequence by ChatGPT

KO

You have four actions: pick up something, keep grasping something, put a thing you have, and move from place A to place B. There is a box in front of you. If someone told you to "bring a box," what kind of action would you take?

If someone told me to "bring a box," I would perform the following actions:

1. Pick up the box: I would bend down, grasp the box, and lift it off the ground.
2. Keep grasping the box: Once I have the box in my hands, I would continue to hold onto it securely to ensure it doesn't fall.
3. Move from place A to place B: I would then carry the box and transport it from its current location (place A) to the desired destination (place B).
4. Put the box down: Finally, upon reaching place B, I would gently place the box down, completing the task of bringing the box.

So, the actions involved in "bringing a box" would include picking up the box, keeping a firm grasp on it, moving it to the specified location, and then putting it down in that new location.

Is planning problem solved?

KO

You have four actions: pick up something, keep grasping something, put a thing you have, and move from place A to place B. There is a box in front of you. There is a cup of coffee on the box. If someone told you to "bring a box," what kind of action would you take?

Action and commonsense in real-world

◆**We don’t verbalize the commonsense if we do not aware of them**

- E.g., no one will say that “Coffee spills when an object with coffee on top is lifted.”

◆**However, what has been clogging up the robot's agenda so far is this kind of common knowledge in the real world**

◆**LLM is pushing the research forward**

- Last five-centimeter problem
- Knowledge and things should be connected

Can LLM be enhanced with prompt?

< 3 / 3 >

KO

You have four actions: pick up something, keep grasping something, put a thing you have, and move from place A to place B. There is a box in front of you. There is a cup of coffee on the box. If someone told you to "bring a box," what kind of action would you take? Note that, a cup of coffee on something will spill if the thing is picked up. To prevent such kind of problem, you need to You also need to think about the motion of lowering the coffee cup to another location before picking the object up. You don't have an action to lower the coffee.

< 2 / 2 >

Given the constraints you've provided, if someone told me to "bring a box" and there is a cup of coffee on the box that will spill if picked up, I would need to take additional steps to prevent the coffee from spilling. Since I don't have an action to lower the coffee, I would need to communicate with the person requesting me to bring the box and clarify the situation.

I would ask if it's acceptable to move the cup of coffee to a safe location before picking up the box, as the current set of actions does not allow for simultaneously carrying the cup of coffee without spilling it while picking up the box.

1. Put down the cup of coffee: First, I would gently place the cup of coffee in a safe location to prevent it from spilling.
2. Pick up the box: After ensuring the cup of coffee is safely placed, I would then pick up the box as previously described.
3. Keep grasping the box: While holding the box securely, I would ensure that I maintain a firm grasp on it.
4. Move from place A to place B: I would carry the box to the desired destination (place B).
5. Put the box down: Finally, upon reaching place B, I would gently place the box down.

Attention depending on task

◆**Which information we are focusing on to do tasks?**

- This question is still not obvious
- **LLM often requires prompt for proper focusing**
  - Is some case, it still fails…

◆**Pretrain model for multimodal models still has a variety**

- This is because the “training objective” enforce what the model can do (and what the model cannot do)
  - E.g., CLIP, EfficientNet

In contrast, the objective of language model is very strong because our language and their meaning have high correlation with distributional hypothesis

What can we do by language?

◆**We can describe a diverse input by language**

- Multimodality, necessary for robot operation
- Bridging with language = bridging with knowledges, physicality

◆**Using large-scale pretrained language model**

- We can finetune a system with small amount of data
- Is there any limitation of pretrained models?

Next generation langrobo

◆Language & robotics research is empowered by LLMs

- However, the problem setting is still setting up a sequence of motion primitives
- Is it able to directly generate motions from language?

◆Language to motion (action to language)

- Recognition is easier than generation
- Some points in motion or action is **not clearly mentioned in language**
  - E.g., speech recognition and speech synthesis
  - Diffusion model? (Motion Diffusion)

Toward building robots with language interface

◆**Understanding both clarified/not-clarified part**

- Understanding the indirect meaning, potential meaning
- Using commonsense
- Using surrounding situation
- In language2motion,
  - how does the system decide not-mentioned points
  - How does the system fit the language to robot's physical limitation

◆**Limitation in the real-world**

- Simulation is not always enough
- We don't have enough data in real-world
- How do we define the necessary information for a task?