Chat with the Environment: Interactive Multimodal Perception Using Large Language Models


[Xufeng Zhao]


Mengdi Li


Cornelius Weber


Muhammad Burhan Hafez


Stefan Wermter

University of Hamburg
Department of Informatics
Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

The Next Generation of Robotics

How do humans/robots perceive the surroundings to uncover latent properties? [1]

- **Common sense**
- **Established knowledge**

• Passive perceptions

Visual monitoring
Auditory monitoring
…

(Epistemic) Uncertainty

Information insufficiency in modalities
Ambiguity in human instructions
Low-resolution sensing
…

• Active perceptions

Weigh
Knock on
Touch
…

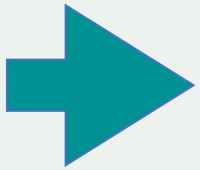
Bridge the gap with LLMs

Robots with hand-crafted design

- Increased complexity
- Difficulties in generalizability and robustness in dynamically changing environments

Humans

- **Common sense**
- **Established knowledge**



Matcha* agent

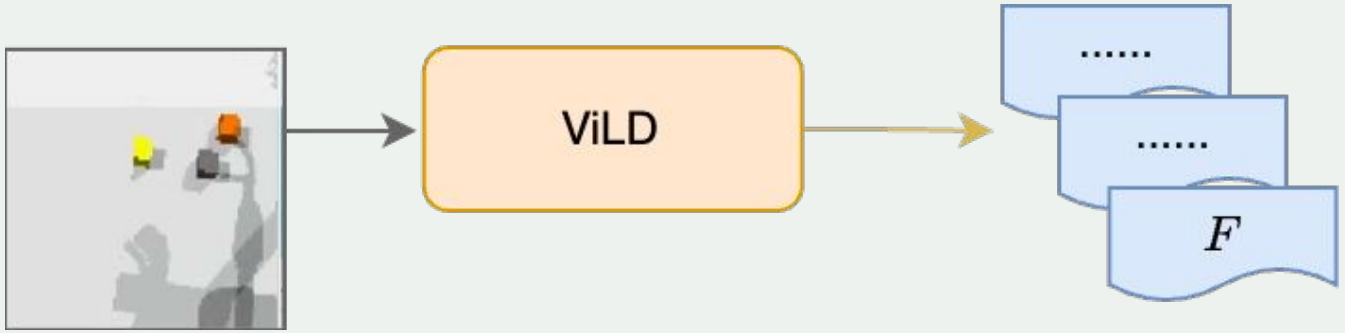
(**M**ultimod**a**l environmen**t** **ch**atting **a**gent)

Robots with LLMs

- **Reasoning / Planning ability with distilled human knowledge inside**
- **In-context learning ability with few-shot prompts**

* By the name of a type of East Asian green tea. To fully appreciate matcha, one must engage multiple senses to perceive its appearance, aroma, taste, texture, and other sensory nuances.

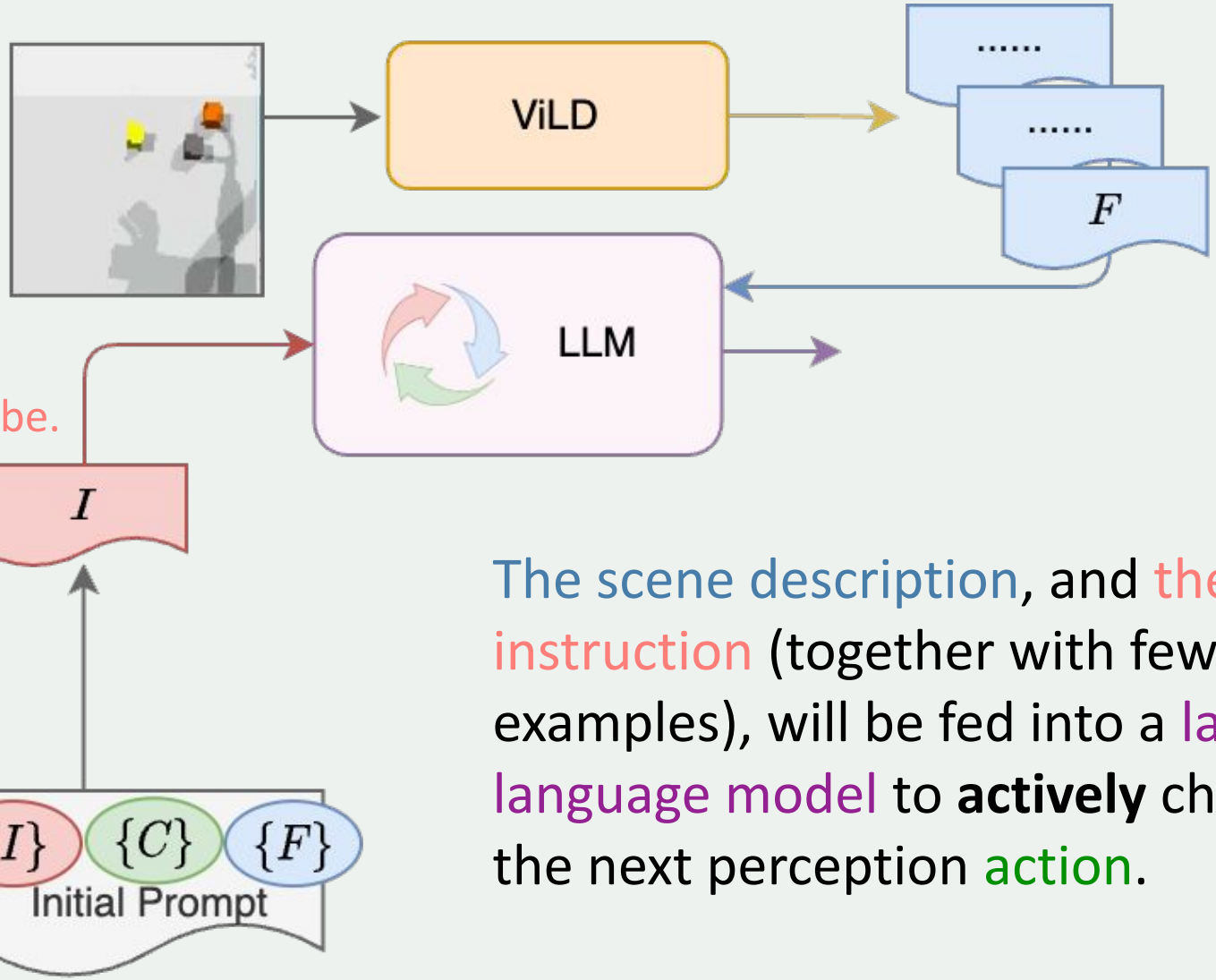
Matcha agent
(Structure)



Yellow block
Orange block
Gray block
…

Start with a vision module to describe the scene

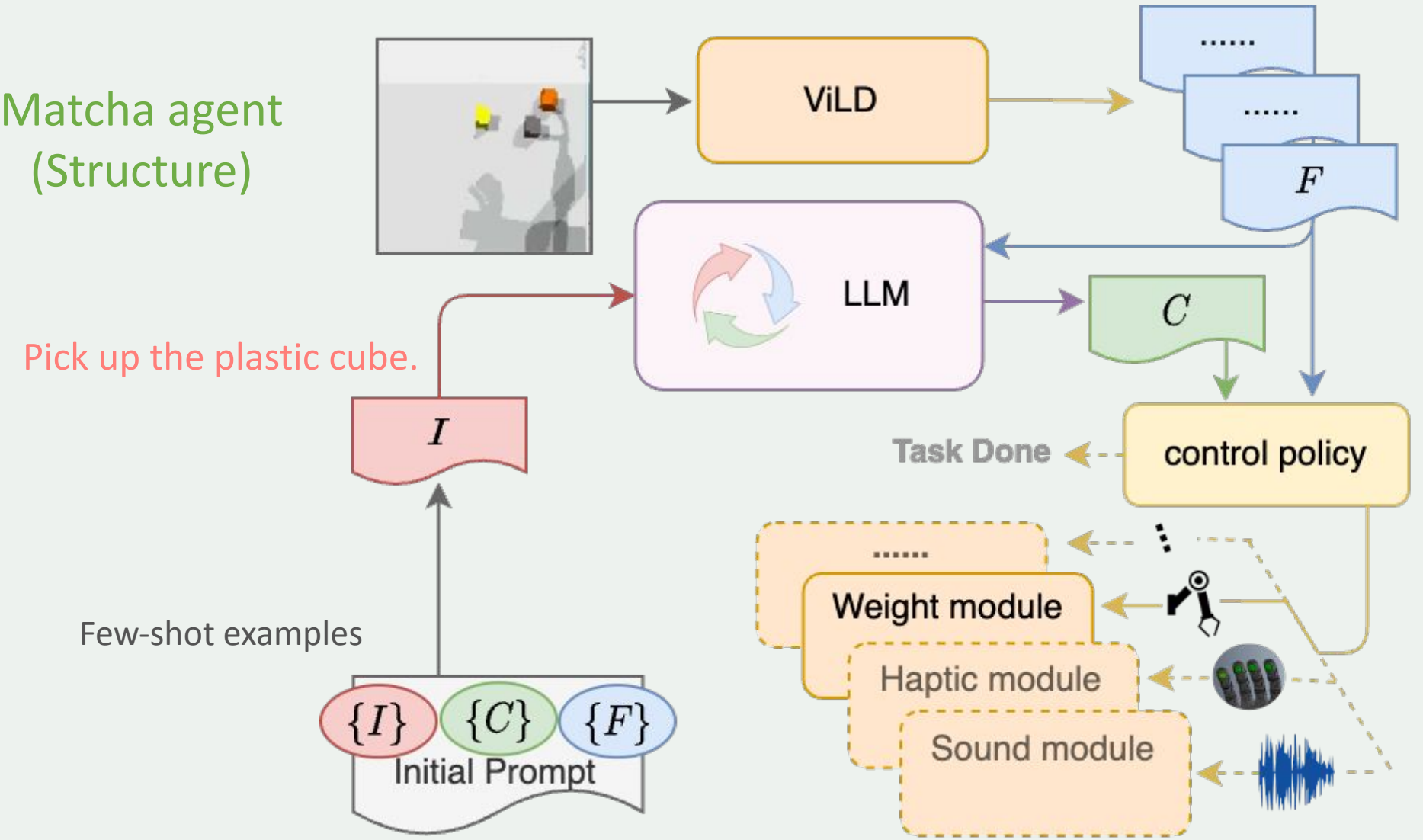
Matcha agent
(Structure)



Yellow block
Orange block
Gray block
…

The scene description, and the task instruction (together with few-shot examples), will be fed into a large language model to **actively** choose the next perception action.

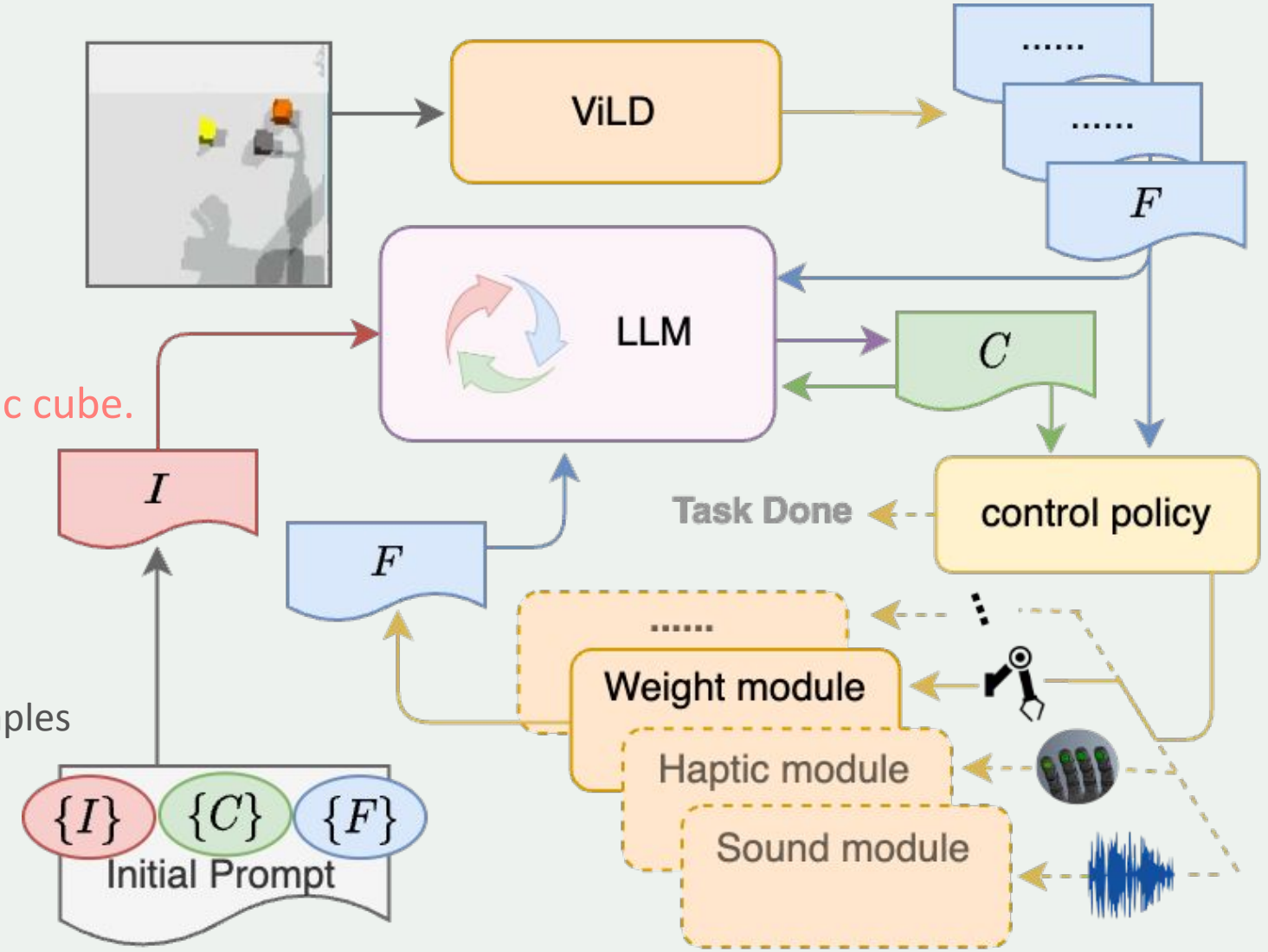
Matcha agent (Structure)



Yellow block
Orange block
Gray block
…

The chosen action will be carried out with motion planning.

Matcha agent (Structure)



Yellow block
Orange block
Gray block
…

Feeding back the multimodal response to the LLM and loop until the task is done.

- Sound module
- Haptic module
- Weight module

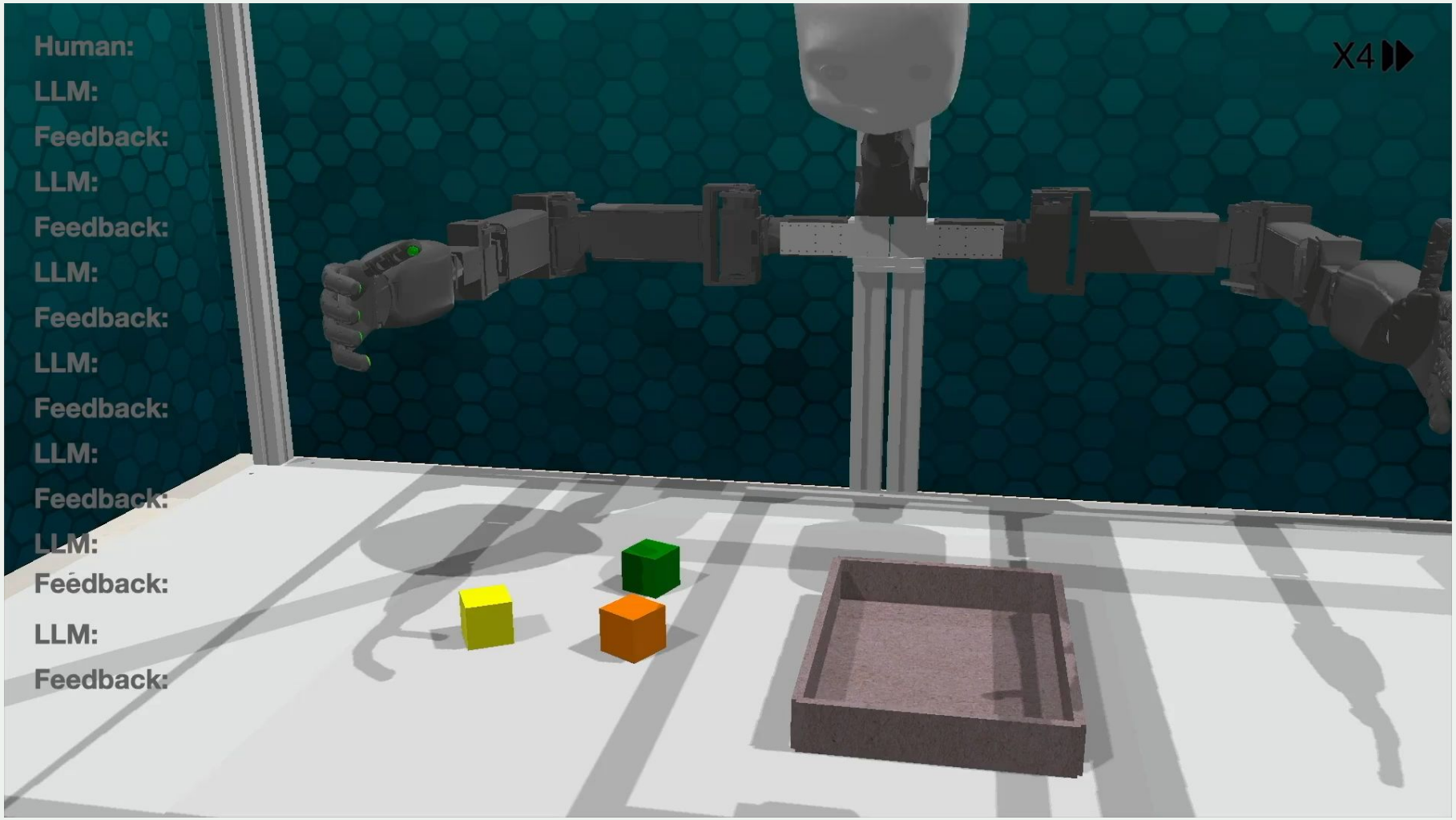
- (Overlap) Similar modality descriptions for **different** materials

- (Conflict) Quite different descriptions for the **same** material

Materials	Impact Sound	Haptics	Weight
Metal	"resonant and echoing", "metallic", "ringing"	"hard and cold", "rigid, cold, and smooth"	"heavy", "300g"
Glass	"tinkling", "tinkling and brittle"	"hard", "hard and smooth", "cold and smooth"	"a little bit heavy", "150g"
Ceramic	"clinking and rattling", "rattling", "tinkling and brittle"	"hard", "tough"	"average weight", "not too light nor not too heavy", "100g"
Plastic	"dull", "muffled"	"hard", "soft"	"light", "30g"
Fibre	"muted", "silent"	"soft", "flexible"	"lightweight", "underweight", "10g"

Matcha agent (In simulation)

- NICOL robot [2]
- Coppeliasim simulator
- LLM: OpenAI API text-davinci-003
- Speed x4



https://youtu.be/rMMeMTWmT0k

[2] Kerzel, Matthias, et al. "NICOL: A Neuro-inspired Collaborative Semi-humanoid Robot that Bridges Social Interaction and Reliable Manipulation." arXiv preprint arXiv:2305.08528 (2023).

Experiment results

LLM	Type of Description	Success Rate
text-ada-001	Indistinct	19.05%
	Distinct	28.57%
text-davinci-003	Indistinct	56.67%
	Distinct	90.57%

*Random guess in principle: 33.33%

- Works without any fine-tuning
- The language instructions can be flexible
- Only a larger language model with strong multistep reasoning ability helps

Generalization, Limitation and Future Work

- No need for massive dataset/interactions to learn the common sense
- Limitations in interpreting the real, **complex**, **dynamic** world with language
  - Large multimodal models
  - Advanced reasoning techniques to decompose tasks
  - …
- Future work: large multimodal models and real-world robots

Thank You for Your Attention!





Xufeng Zhao
Mengdi Li
Cornelius Weber
Muhammad Burhan Hafez
Stefan Wermter

University of Hamburg
Department of Informatics
Knowledge Technology

Matcha

Logos

OCTOBER 1 - 5, 2023

IEEE/RSJ International Conference
on Intelligent Robots and Systems

DETROIT

IEEE/RSJ International Conference
on Intelligent Robots and Systems
OCTOBER 1-5, 2023

- Sound classification accuracy: 93.33%
- The robot can randomly knock on an object among three, and classify the material until the one that is classified as the target. In theory, the success rate is computed as 1/3 p + 2/3 p^2 = 89.18%.