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

Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, Stefan Wermter

Knowledge Technology, University of Hamburg

Matcha

We propose the Matcha **framework,** comprising an LLM and multiple **multimodal modules**, enabling the robot to engage with its surroundings through high-level **LLM planning**.

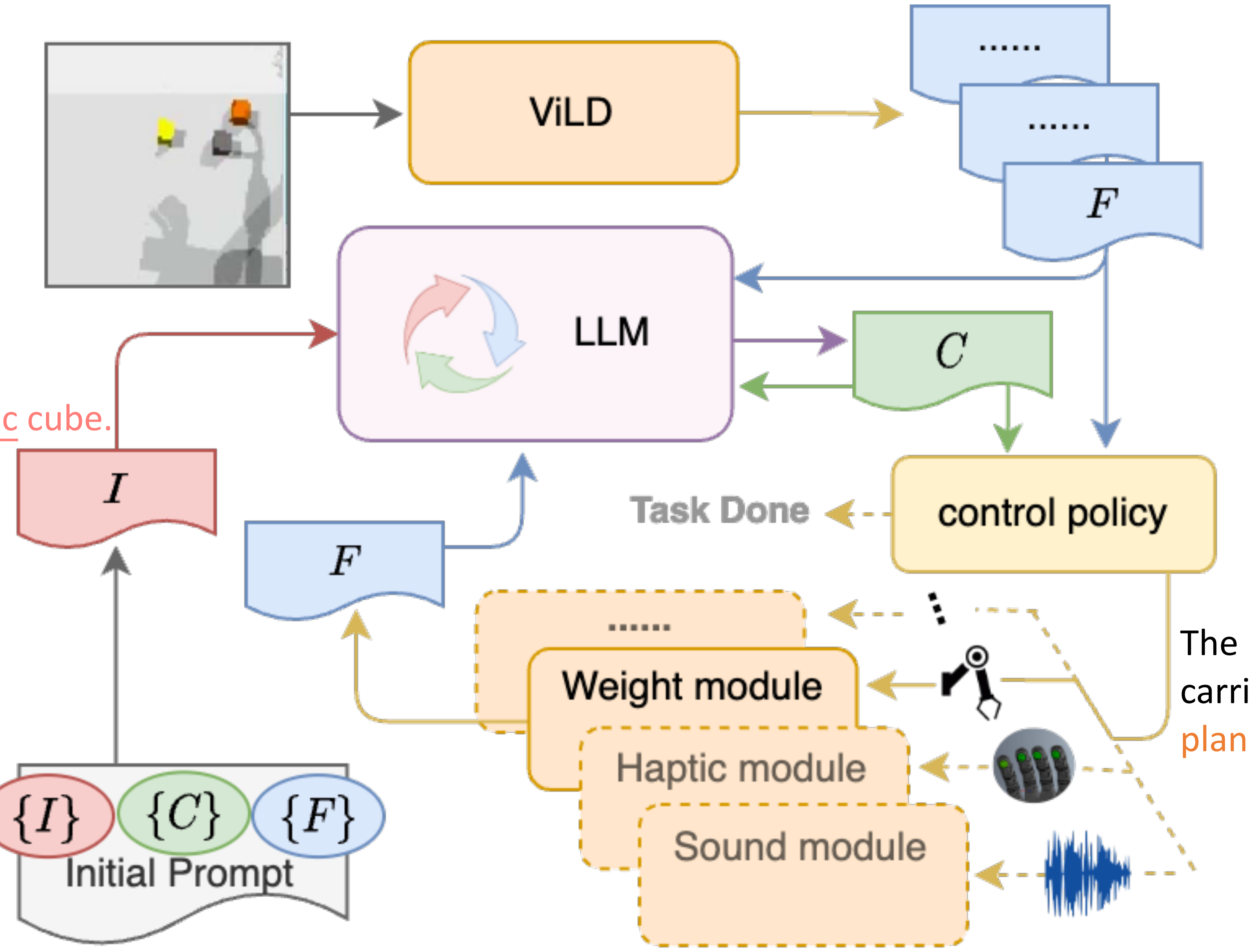
(Multimodal environment chatting agent)

Robotic Perceptions

- **Passive perceptions**
  - **Epistemic uncertainty**
- **Active perceptions**
  - **Increased complexity**
  - **Generalizability**
- **Robots with LLMs**
  - **Causal reasoning ability with distilled human knowledge inside**
  - **In-context learning ability with few-shot prompts**

Matcha Architecture

Start with scene description (with vision module)



Yellow block
Orange block
Gray block
...

Task input:
Pick up the plastic cube.

Few-shot examples
I: instruction
C: commanding action
F: Environment Feedback

The chosen action will be carried out with motion planning.

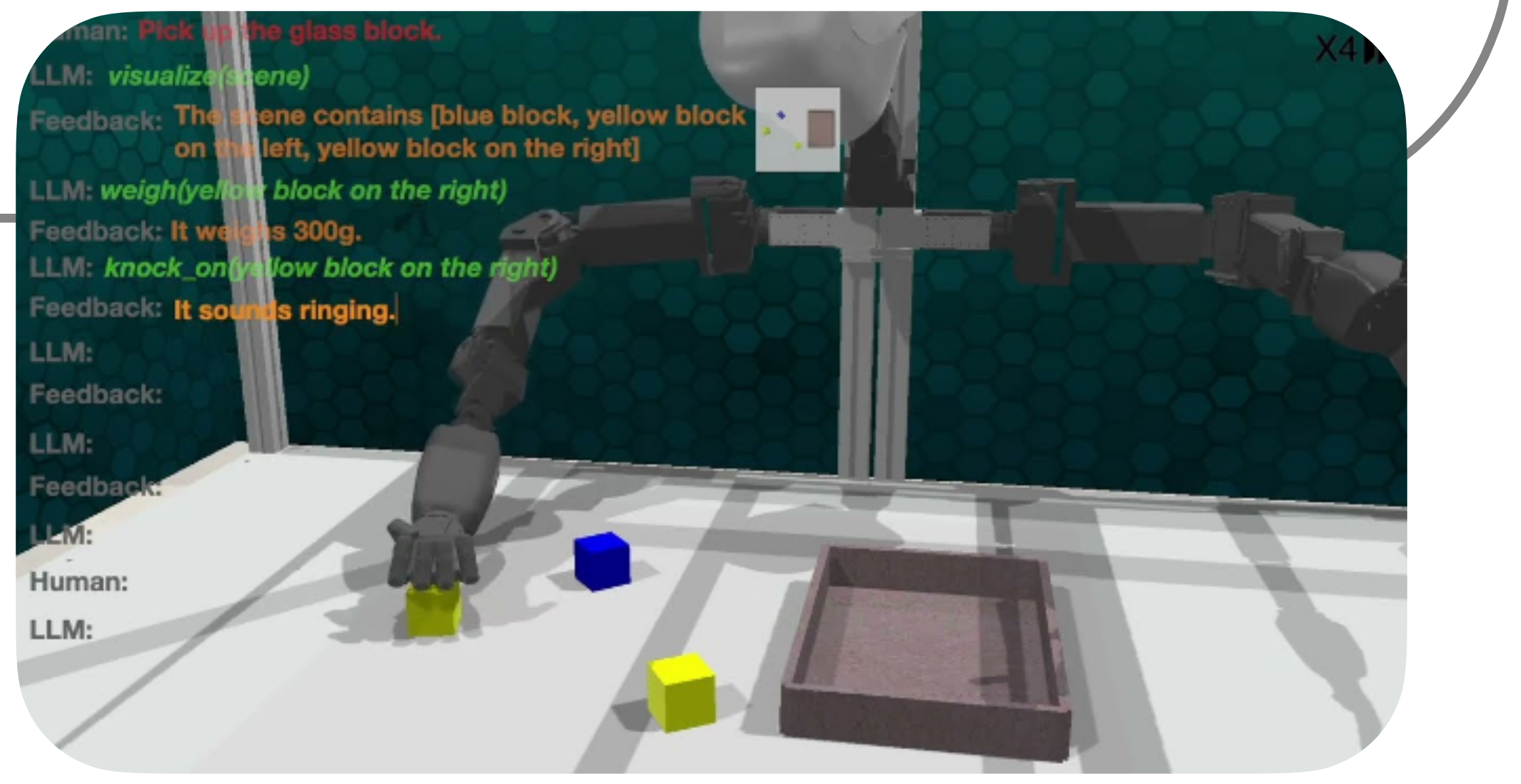
Feeding back the multimodal response to the LLM and loop until the task is done.

- Sound module
  - Similar modality descriptions for different materials
  - Quite different descriptions for the same material
- Haptic module
- Weight module

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%

- NICOL robot
- Coppeliasim simulator
- LLM: OpenAI API text-davinci-003
- Works without any fine-tuning



Human: Pick up the glass block.
LLM: visualize(scene)
Feedback: The scene contains [blue block, yellow block on the left, yellow block on the right]
LLM: weigh(yellow block on the right)
Feedback: It weighs 300g.
LLM: knock_on(yellow block on the right)
Feedback: It sounds ringing.
LLM:
Feedback:
LLM:
Feedback:
LLM:
Human:
LLM: