

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

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

Abstract—Programming robot behavior in a complex world faces challenges on multiple levels, from dextrous low-level skills to high-level planning and reasoning. Recent pre-trained Large Language Models (LLMs) have shown remarkable reasoning ability in few-shot robotic planning. However, it remains challenging to ground LLMs in multimodal sensory input and continuous action output, while enabling a robot to interact with its environment and acquire novel information as its policies unfold. We develop a robot interaction scenario with a partially observable state, which necessitates a robot to decide on a range of epistemic actions in order to sample sensory information among multiple modalities, before being able to execute the task correctly. *Matcha* (Multimodal environment chatting) agent, an interactive perception framework, is therefore proposed with an LLM as its backbone, whose ability is exploited to instruct epistemic actions and to reason over the resulting multimodal sensations (vision, sound, haptics, proprioception), as well as to plan an entire task execution based on the interactively acquired information. Our study demonstrates that LLMs can provide high-level planning and reasoning skills and control interactive robot behavior in a multimodal environment, while multimodal modules with the context of the environmental state help ground the LLMs and extend their processing ability. The project website can be found at <https://matcha-agent.github.io>.

I. INTRODUCTION

How do humans perceive the surroundings to uncover latent properties?

Suppose you are presented with an uncommon object in a strange shape and of unknown material, you may explore its properties in both passive and active ways, if possible, e.g. by observing the geometry, touching and even knocking on the surface in order to deduce its exact functionalities from the feedback. Unnecessary explorations, which could be essential for other scenarios such as smelling, will not be performed in this context unless something counterintuitive happens. We humans naturally perform these **multimodal observations and examinations** in daily life through **common sense and established knowledge**, and over time we adapt with the accumulation of experience [4].

This is also now possible for a robot well-equipped with multiple sensors and Large Language Models (LLMs). An environment may be filled with rich content, and the robot can be overwhelmed with diversified sensory stimuli. An

This research was funded by the German Research Foundation (DFG) in the project Crossmodal Learning (TRR-169) and the China Scholarship Council (CSC).

The authors are with the Knowledge Technology Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany. E-mail: {xufeng.zhao, cornelius.weber, burhan.hafez, stefan.wermter}@uni-hamburg.de, mengdi.li@studium.uni-hamburg.de.

*Corresponding author.

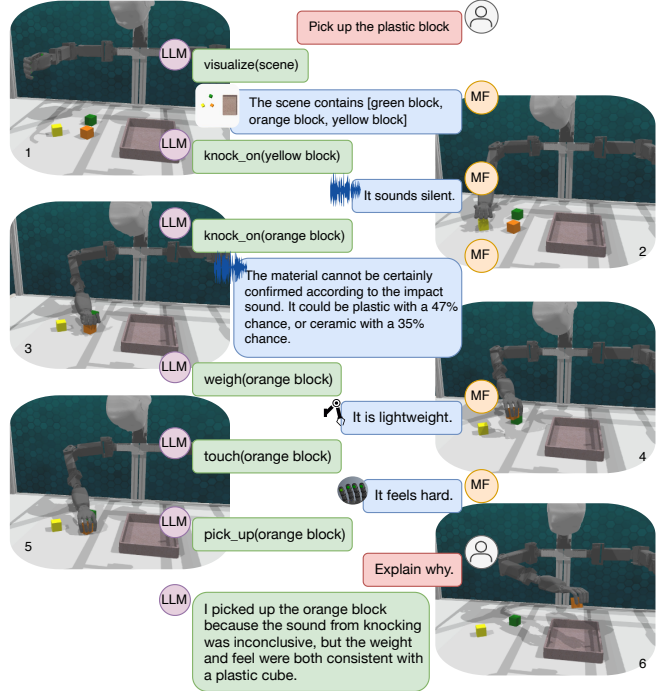


Fig. 1: Given instruction from a human, the robot recurrently “chats” with the environment to obtain sufficient information for achieving the task. An LLM generates action commands to interactively perceive the environment and, in response, the environment provides multimodal feedback (MF) through multimodal perception modules.

intelligent robot should 1) wisely choose stimuli to attend to, avoiding eagerly being bogged down into details, and 2) respond accordingly to the resulting sensations in the context of a specific task.

A. Interactive Multimodal Perceptions

Like humans, robots can perceive the environment in either a passive or an interactive way [12]. *Passive perception* refers to ways such as visual or auditory monitoring, and it allows robots to quickly acquire information without intervening with the surroundings. However, the passive manner has its limits, among which the most outstanding one is its impotency when facing *epistemic uncertainty* [6], the uncertainty because of lacking knowledge.

Epistemic uncertainty inevitably arises from diverse sources, e.g. from the ambiguity in human instructions, from low-resolution sensing (e.g. reduced image size for convolution), or from insufficient modalities. Many of them

can only be reduced with *interactive perception*, in which a robot actively interrogates the environment to increase accuracy and even uncover latent information. For example, when being asked to deliver a *steel* screw instead of one with a similar color & shape but made of *aluminum*, an assistant robot may need to locate possible candidates with *passive* vision and further, *interactively*, resort to a weighing or a magnetic module for confirmation.

Despite the promising advantages, interactive perception is less common than the passive manner because it entails increased complexity [14]. Efforts are needed to design a mediating system to handle various sensory data and to adapt to changes in the conditions of both the robot and the environment, such as a new robotic modular being available or the involvement of novel objects.

B. Chatting with the Environment

LLMs have been showing incredible potential in areas besides robotics [1], [7], [15], [16]. Human knowledge that resides in LLMs can help a robot abstract and select only suitable features, e.g. relevant to the region of interest or informative modalities, to simplify the learning process. Moreover, in terms of generalizability, the knowledge of LLMs allows a behavioral agent to adapt efficiently to novel concepts and environmental structures. For instance, when being asked to *use one adjective for each to describe how a sponge and a brick feel*, ChatGPT¹ will respond with “soft” and “hard” respectively. This is helpful for a robot with a haptics sensing module to distinguish between these two novel, never-seen objects.

LLMs are usually generative models that predict tokens to come, but with certain designs, e.g. conversational prompts, LLMs are capable of generating chat-fashion texts. This allows their integration with a robot to not only plan with respect to a robot’s built-in ability [27], [1] but also respond according to environmental feedback.

However, they cannot directly process application-specified raw multimodal data. We resort to modular perceptions for each modality that are separately trained before being plugged into the LLM backbone. Each module semantically translates the resulting multimodal sensations into natural language that can be understood by LLMs and processed in a unified manner.

Our contributions are threefold. Firstly, we establish a manipulation scenario with multimodal sensory data and language descriptions. Secondly, we propose **Matcha**² (**M**ultimodal **e**nvironment **c**hatting) **a**gent, where an LLM is prompted to work in a chatting fashion, thus having continuous access to environmental feedback for contextual reasoning and planning. Finally, we show that LLMs can be utilized to perform interactive multimodal perception and behavior explanation. Accordingly, an interactive robot can make reasonable and robust decisions by resorting to LLMs to

examine objects and clarify their properties that are essential to completing the task (see Fig. 1).

II. RELATED WORK

Multimodal Learning and Robotic Information Gathering. Research in multimodality in robotics nowadays attracts growing attention [2] because of its success in, for example, audio-visual learning [28], [26], [30] and language-visual learning [20], [21]. It is beneficial and sometimes essential for a robot to learn from multimodality because one modality could carry some distinct information, e.g. tones in speech, that cannot be deduced from another. [13].

Capable robots require managing one or several sensors to maximize the information needed for disambiguation [4] regarding a specific goal. This problem is known as *active information acquisition* [3], [24] or, particularly in robotics, *robotic information gathering* [19], where robots have to properly select perceiving actions to reduce ambiguity or uncertainty. Besides handcrafted rules, some information advantage measures, e.g. entropy or information gain, are usually employed to maximize [3]. However, the combination of multimodal data is usually challenging. There are studies on fusing multimodal data according to their uncertainties, but this may face numerical instability and is difficult to transfer from one application to another [25]. Instead of directly fusing the multisensory data in a numerical space, we propose to use multimodal modules to translate them into natural language expressions that an LLM can easily digest.

Large Language Models in Robotic Planning. Very recent works use LLMs to decompose high-level instructions into actionable low-level commands for zero-shot planning. They use LLMs as a planner to autoregressively select actions that are appropriate with respect to the instruction according to application-based prompts [27], the semantic similarity between mapped pairs [10], or the contextual language score grounded on realistic robot affordances [1]. Other approaches ground LLM knowledge in human interaction [7] or many other various fields where domain knowledge is distinct and modular frameworks can be composed via language as the intermediate representation [18], [16], [27].

However, these works design a robot to form a planning strategy with *built-in knowledge*, rather than *interact* with the surroundings and make decisions based on *actively collected information* from the environment. There is no feedback loop for their LLMs to perceive the environmental cues, such that only “blind” decisions are made in the robotic unrolling process. In contrast, our interactive architecture allows LLMs to access the environment state from multiple modalities for adaptive planning.

III. METHODOLOGY

A. Architecture

We propose **Matcha** (**m**ultimodal **e**nvironment **c**hatting) **a**gent which is able to interactively perceive (“chat” with) the environment through multimodal perception when the information from passive visual perception is insufficient for completing an instructed task. The epistemic actions are

¹<https://openai.com/blog/chatgpt/>

²By 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.

executed autoregressively until the agent is confident enough about the information sufficiency in that situation.

Fig. 2 provides an overview of the architecture of Matcha agent. It is a modular framework of three parts: an LLM backbone, multimodal perception modules, and a low-level command execution policy. They connect to each other with language as the intermediate representation for information exchange.

To be specific, given a high-level instruction, especially the one that Matcha cannot directly perform with the command policy alone, the LLM backbone will reason the situations and select the most contextually admissible perceiving command to gather information. After the execution of the policy module, the resulting environmental response is processed by a correspondingly evoked multimodal perception module into semantic descriptions, e.g. “clinking sound” by an auditory module after the “knock on” action. Finally, the executed command itself as well as the environmental state description are fed back to the LLM for future planning.

The LLM is used in a few-shot manner without any need for fine-tuning, being independent of other components. Policy and perception modules can be separately designed and plugged into the framework whenever needed. Intrinsically linked by natural language, this framework is flexible and can scale and adapt easily to possible robotic upgrades or diverse robotic scenarios.

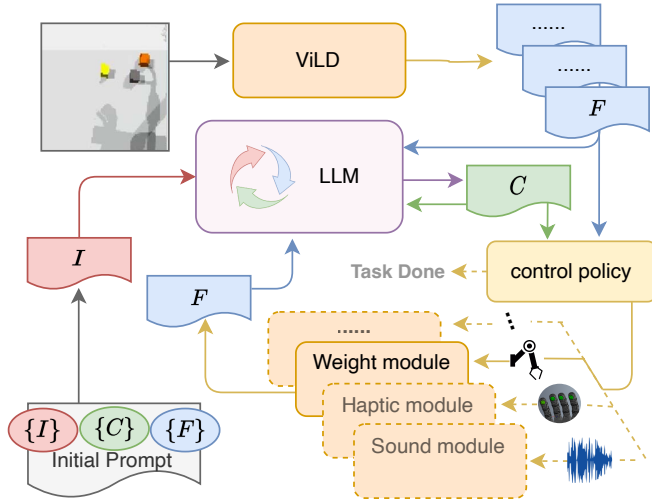


Fig. 2: Overview of Matcha. The framework contains an LLM, multimodal perception modules, and a language-conditioned control policy. These components communicate with each other with natural language as the intermediate representation. Three types of language information are involved in composing the prompt: I is a language instruction from the user, C is a language command produced by the LLM, and F is semantic feedback from multimodal perceptions. Dotted lines indicate possibly evoked paths.

B. Multimodal Perception and Execution Policy

We select a commonly feasible suit of modalities and a language-conditioned policy as an example implementation

of our framework. Other varieties for specific scenarios can also be easily integrated due to the flexibility of modularity of the framework. Detailed experimental implementations will be introduced in Sec. IV.

1) *Vision*: Usually, a monitoring camera is the cheapest option for a robot to passively perceive such rich information. We employ pre-trained ViLD [9], an open-vocabulary visual detection model, as the vision perception module to detect objects with their categories and positions in the scene. Then, the results will be delivered to a policy module for identification and execution. Meanwhile, a prompt template “The scene contains [OBJ1, OBJ2, ...]” is applied to construct a scene description, which enables the LLM to have an initial impression of the environment. Typically, pre-trained vision models are not designed to discern attributes that extend beyond those easily extractable from topology or textures, such as material composition. The use of low-resolution images for expedited processing exacerbates the loss of information concerning such attributes. In our experimental approach, we prioritize demonstrating the integration of diverse modalities instead of extensively fine-tuning ViLD to encompass all aspects.

2) *Impact Sound*: Impact sound commonly occurs from time to time, and can be useful for robotic multimodal learning [28]. Though it can be passively collected with a microphone attached to the robotic end-effector, without intentional intervention by the robot, a “knock on” action in our case, a microphone may only be able to collect background noise. This auditory perception module classifies the consequent impact sound into a description and then wraps it in a natural language form. Actually, a clip of audio may contain sufficient information for some of the usage, e.g. to distinguish metal from glass [8]. However, it may not be the case for other scenarios, for example, to select the only targeted one among a set of similar “dull” sounds that could indicate either plastic, wood or hard paper. Therefore, we showcase both of the designs, i.e. one with a specific material classification (e.g. “glass”) and another with solely low-level and non-distinct descriptions (e.g. “tinkling”). The modular output is also wrapped with templates to a full sentence such as “It sounds tinkling”, to guarantee processing consistency with LLMs.

3) *Weight*: Weight measurements are usually obtained via the torque exerted on the robotic arm subsequent to the execution of an “weighing” action. The weight information is directly translated into natural language like “It is lightweight” or “It weighs 30g”. Note that with implicit clarification of the scenario and the type of objects that a robot is manipulating, LLMs can interpret numerical values into contextual meanings.

4) *Haptics*: Haptic perception is extremely important for humans to interact with their surroundings. It also provides a potential for robots when acquiring information related to physical properties, including hardness, texture, and so on. However, a high-resolution tactile sensor is costly and not worthwhile for many applications. Therefore, in our case, we only use highly abstract descriptions for the force-torque

feedback subsequent to a “touching” action on an object, e.g. “It feels soft” or “It feels hard and smooth”.

5) *Execution Policy*: The execution policy is conditioned on the generated command by an LLM and the visual information provided by the vision perception module. Once an actionable command together with an identified target is suggested by the LLM, the policy module locates the targeted object and executes a certain action. Meanwhile, the environmental feedback will be concurrently collected for multimodal perception modules for further post-processing as demonstrated above.

C. Prompt Engineering

An issue of grounding LLMs on robotic scenarios is that some of the suggestions generated by LLMs are not executable for a specific robot [1], [10], which stems from the fact that LLMs are pre-trained with extremely large open-domain corpora, while the robot is constrained by its physical capability and application scenarios, e.g. a tabletop robot is not able to perform “walk” action.

In this work, the LLM is applied for few-shot prompting [16], [27], in which all the executable commands are defined together with several task examples as the initial “chat” history. See Tab. I for the leading prompt which enables the LLM to ground on the specific scenario and follow the contextual patterns for commanding the execution policy.

TABLE I: The snippet of the 5-shot prompt setting. The other four exemplars are omitted here due to the content limit.

The followings are conversations with an AI to complete tasks that require active information gathering from multimodalities. Otherwise, the materials of objects are unknown, and it will be ambiguous for an AI to choose the right object. AI has the following skills to help complete a task:

1. “robot.knock_on()”: to knock on any object and hear the sound to determine the material it consists of. Most of the materials can be determined by this skill.
2. “robot.touch()”: to touch with haptics sensors. It is useful for some of the materials.
3. “robot.weigh()”: to weigh objects if the knocking method is not proper.
4. “robot.pick_up()”: to pick up the targeted object. After this skill is performed, the episode will terminate with the result.

Note that the tasks are always set to be accomplishable, and the selected skill should start with a “>” symbol.

...

Human: “pick up the glass block” in the scene contains [yellow block, blue block, green block]
 AI: robot.weigh(yellow block)
 Feedback: It weighs light.
 AI: robot.weigh(blue block)
 Feedback: It weighs a little bit heavy.
 AI: robot.knock_on(blue block)
 Feedback: It sounds tinkling.
 AI: robot.pick_up(blue block)
 done()
 ...

We found that only language models that are large enough can follow the patterns in the prompt strictly, i.e. only generate commands that have been defined in strictly case-sensitive letters and with the same amount of allowed parameters for each, while small ones can hardly obey this constraint and generate unexpected commands, which brings extra demands

for tuning. As the action planning is performed by LLMs constrained by a given prompt, the proposed framework demonstrates high flexibility and generalizability upon the possible incorporation of novel actions or perception modules into the system.

IV. EXPERIMENTS

A. Experimental Setup

We evaluate the proposed framework in an object-picking task: a robot is instructed to pick up an object that is referred to by a latent property – *material* – which is, however, not visually distinguishable under our settings. Tasks are intentionally designed such that information from a single modality could be insufficient to determine object properties, while other perception sources can provide compensations to reduce or eliminate this ambiguity. For example, glass and metal surfaces could exhibit similar hard and smooth properties upon contact, in which case differences in impact sound can aid in further differentiation. Tab. II lists variational multimodal descriptions of the materials. These properties are wrapped as natural language sentences before being fed back to the LLM.

Experiments are done in CoppeliaSim³ simulations with the NICOL robot [11], where several blocks in various colors, materials, weights, and surface textures are randomly selected and placed on the table next to a brown container (see Fig. 1). The ViLD [9] model is meant to be easily generalized to describe complex scenes despite the simplicity of the object setting here. After detection, the objects are represented universally by their name, which serves as a parameter for the action function to identify. Objects with the same color will be distinguished as “.. on the left/right” given the simplicity of avoiding more than two duplicated colors for the same shape. The desktop robot is equipped with two *Open-Manipulator-Pro* arms⁴, but only its right arm is activated to operate. It is capable of executing actions in [“knock on”, “touch”, “weigh”, “pick up”], with a parameter to indicate the targeted object. The first three actions correspond to the interactive perception of impact sound, haptics, and weight respectively, and the last action finalizes the task by picking and transporting an object into the box. Each instruction is guaranteed to be achievable with the capability of the robot.

Due to the lack of support for physics-driven sound and deformable object simulation in Coppeliassim, we have implemented reasonable alternatives. For the haptics of objects, we simplify haptic perception by assigning variational descriptions regarding its material, e.g. fibrous objects are usually perceived as “soft” and a plastic object can be either “soft” or “hard”. Note that advanced implementations can also be achieved using a neural network as is used in the sound perception module when haptics data for deformable objects is available. For the impact sound, we split the YCB-impact-sound dataset [8] into training and testing sets and

³<https://www.coppeliarobotics.com/>

⁴https://emmanual.robotis.com/docs/en/platform/openmanipulator_p/overview/

augment them with tricks such as shifting, random cropping, and adding noise. The training set is used to train our auditory classification neural networks, while the audios in the testing part are randomly loaded as an alternative to run-time impact sound simulation for the materials mentioned,

Sound can be informative, though not perfect, for determining materials [8]. Besides showing the mediating ability of multiple modalities by the LLM, we further investigate its reasoning ability by employing indistinct descriptions instead of exact material labels.

- *Distinct description*: the sound module describes sound feedback by the corresponding material name and its certainty from the classification model, e.g. “It is probably glass” or “It could be plastic with a 47% chance, or ceramic with a 35% chance”. The distinct description setting is more task-oriented, and it examines the robot’s ability to mediate multiple sensory data for disambiguation.
- *Indistinct description*: we listed some commonly used indistinct sound descriptions in human communications in Tab. II, e.g. using “dull” to describe the sound from a plastic block and “tinkling” to describe the sound for both ceramic and glass objects. This setting is more task-agnostic and thus has the potential for generalization. Moreover, it compels the LLM to infer “professional” material terminology from ambiguous yet multimodal descriptions.

The online OpenAI *text-davinci-003* API⁵ is applied as the LLM backend because it demonstrates robust instruction-following ability and outstanding reasoning performance.⁶ We also evaluate with a weaker but far less expensive LLM *text-ada-001*, a GPT-3 model which is usually fast and capable of simple tasks, under the same setting as comparison.

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”

TABLE II: Property descriptions of different materials.

⁵ <https://platform.openai.com/docs/models/gpt-3>

⁶The *code-davinci-002* is not chosen because it is common sense instead of the ability of code completion that matters to the active perception. Upon the time that this experiment was carried out, the *text-davinci-003* model is the state-of-the-art GPT-3.5 model available; while the later released ChatGPT or GPT-4 model showcases the impressive improved abilities of reasoning, future works will explore the potential of these models.

B. Results

We test the proposed framework Matcha in 50 randomly generated scenarios for each setting and report the success rate.

We report that the impact sound classification model pre-trained with the selected materials achieves an accuracy of 93.33%. When using distinct descriptions, suppose we are making hard-coded rules to utilize the sound module to identify the targeted material, 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 computes as $\frac{1}{3}p + \frac{2}{3}p^2|_{p=93.33\%} = 89.18\%$, where p is the modular accuracy. Usually, other modalities, in this case, are not as distinct as sound, and it could be non-ideal for humans to craft such fusion rules for a possible slight improvement. Therefore, the theoretical success rate with only the sound module will be used as our baseline for analysis. Note that this is a reasonable rule that humans will follow, thus it can also be regarded as the upper bound for Matcha if it worked with only impact sound.

Unsurprisingly, Matcha achieves a relatively higher success rate of 90.57% compared to the ideal theory baseline, as it utilizes compensatory information from other modalities in addition to sound. When using the indistinct description of impact sound, Matcha is still able to achieve a success rate of 56.67%, which is larger than a chance success rate of 33.33% achieved by randomly picking one from the three. This result is remarkable as it performs few-shot deduction with only indistinct adjectives available. By analyzing the failure cases, we found that the similar descriptions of glass and ceramic in terms of impact sound, haptics, and weight make it challenging to distinguish one material from the other. This is not an issue with distinct descriptions, where the sound classification network directly predicts the material, preventing a significant loss of information in the translation from sensory data to language. While the system displays remarkable reasoning ability, the results still highlight the significance of a well-designed modular description that is not overly abstract in facilitating interactive perceptions.

We observe that only strong LLMs (GPT-3 *text-davinci-003* in our experiments) can perform the task following the primary prompt as we defined, while weak ones possibly generate commands that are not executable and show degraded performance (see Tab. III). For example, GPT-3 *text-ada-001*⁵, a smaller model in the GPT-3 model series, may generate commands like “*robot.knock_on(metal block)*” or “*robot.weigh(yellow block, blue block)*”, which is not feasible for the policy of the robot.

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

TABLE III: Effect of different LLMs on success rate.

C. Case Studies

We provide case studies to showcase the interactive perception ability of Matcha in multimodal environments. Following the convention of this paper, texts containing information from humans, LLMs, and the environment are indicated in red, green, and blue, respectively. Fig. 3 gives a completion

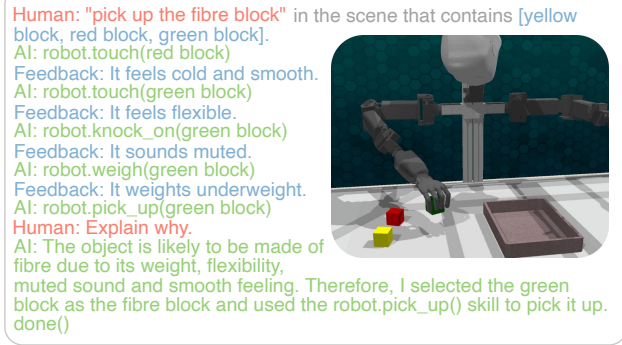


Fig. 3: A successful example in which the robot deduces “fiber” material with indistinct descriptions of impact sound.

that the agent can deduce the latent material property from multimodal yet indistinct descriptions. During the planning process, the agent decidedly terminates exploration of the “red block” and instead engages in interacting with the “green block” exhaustively for a comprehensive examination. Common sense and, moreover, established knowledge in the LLM enables Matcha to efficiently interact with the environment. Matcha’s proficient behaviors provide evidence of effective reasoning, as it aligns with the subsequent explanation provided by the LLM, namely, that fiber can often be considered “flexible” rather than “cold and smooth”. The example depicted in Fig. 4 presents a fascinating observation: the impact sound of the “orange block” suggests it is more likely to be plastic than metal, but Matcha accurately distinguishes it from plastics after engaging in the interactive perception of weight and haptics. This showcases the potential of multimodal perception to improve classification accuracy. Fig. 5 provides a failure case with indistinct descriptions where the robot picks up a ceramic block when asked to pick up the one made of glass. The underlying cause of this failure is the sensing similarity between glass and ceramic, which creates difficulty in resolving epistemic uncertainty.

D. Discussion

Weak LLMs, e.g. ones without fine-tuning on instruction alignment [17], may not have sufficient capability for precise planning, and thus may require carefully engineered prompts or other grounding techniques. On the other hand, strong LLMs exhibit impressive in-context learning abilities. These observations highlight the potential of leveraging knowledge within strong LLMs, as it enables the successful execution of tasks that were previously deemed infeasible. LLMs can derive significant advantages from utilizing common knowledge, being robust to various instructions regardless of their changes in synonym, linguistic structure or even

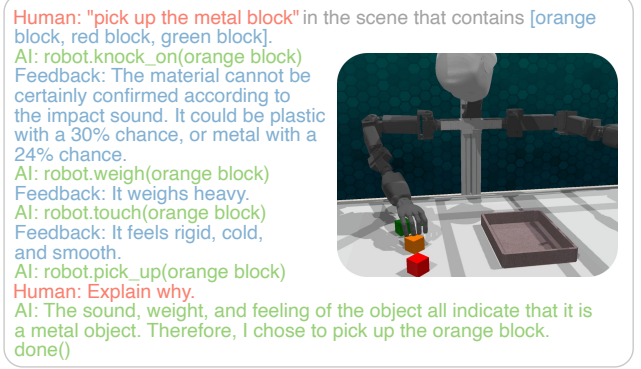


Fig. 4: A successful example with a distinct description of impact sound. This example shows that by leveraging multimodal perception, LLM rectifies the misclassification that may occur when relying solely on sound modules.

semantic meanings out of the scope that the robot is initially designed within, e.g. an instruction variation from “the metal block” to “a block that may be suitable for cracking a nut”, in which the robot has to establish a meaningful connection between the object’s multimodal perceptions and the required utility.

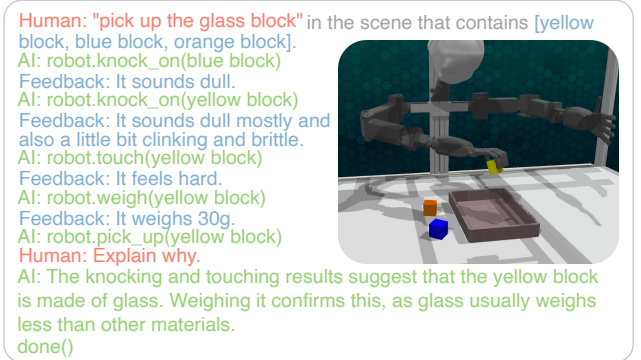


Fig. 5: An example in which the agent fails to distinguish glass and ceramic in the setup of using indistinct descriptions of impact sound.

Nevertheless, the reasoning trace may not always align with human expectations. There are cases that LLMs may prematurely draw conclusions due to their limited logical reasoning ability, particularly when faced with a task that requires reasoning from a long list of facts.

V. GENERALIZATION, LIMITATION AND FUTURE WORK

The Matcha framework exhibits a high degree of generalizability thanks to the commonsense knowledge inside LLMs. Without LLMs, a control algorithm, e.g. one trained with reinforcement learning [14], [22], may require massive datasets/interactions to learn the common sense [22] of collaborating different modalities, yet being less efficient and generalizable.

However, interpreting the real world with language can be limited to the complexity of the task and the environment

dynamics. For example, advanced reasoning techniques such as decomposing may be required to deal with a complicated task, where the task is decomposed into several sub-tasks to tackle separately. This automatic operation highlights the flexibility of LLMs but also poses challenges to the static language expression of a complex world — The vision-to-language module should be called multiple times with flexible queries. This brings the requirement of vision-enabled LLMs [29], [5], built on which the reasoning can be malleable. But multimodal LLMs are yet less controllable and accurate in terms of describing the scene compared with a templated module.

Despite current limitations, multimodal LLMs gain increasing attention due to their great potential and flexibility. Future work will explore the multimodal models [23], [5] to leverage unified features.

VI. CONCLUSIONS

LLMs have shown their impressive ability in language generation and human-like reasoning. Their potential for integration and enhancement with other fields has attracted growing attention from different research areas. In this work, we demonstrate the superiority of using an LLM to realize interactive multimodal perception. We propose **Matcha**, a multimodal interactive agent augmented with LLMs, and evaluate it on the task of uncovering object-latent properties. Experimental results suggest that our agent can perform interactive multimodal perception reasonably by taking advantage of the commonsense knowledge residing in the LLM, being generalizable due to its modularity and flexibility.

While strong LLMs perform well for tasks that require general knowledge, training and maintaining LLMs locally is currently costly, given the large computation and memory resources required by such models. Future works will involve distilling the domain-specific knowledge from LLMs into more manageable local models, which can offer greater flexibility and control while maintaining high levels of performance for robotic applications. Furthermore, there is a necessity for additional investigation of prompt engineering and multimodal LLMs to augment the ability for complex dynamics in the real world.

REFERENCES

- [1] Michael Ahn, Anthony Brohan, Brown, and et. al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [2] Cem Akkus, Luyang Chu, and et. al. Multimodal deep learning. *arXiv preprint arXiv:2301.04856*, 2023.
- [3] Nikolay Asenov Atanasov. *Active Information Acquisition with Mobile Robots*. University of Pennsylvania, 2015.
- [4] Kobus Barnard, Keiji Yanai, and et. al. Cross modal disambiguation. *Toward Category-Level Object Recognition*, 2006.
- [5] Anthony Brohan, Noah Brown, and et. al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv Preprint arXiv:2307.15818*, 2023.
- [6] Carlos Celemin and Jens Kober. Knowledge-and ambiguity-aware robot learning from corrective and evaluative feedback. *Neural Computing and Applications*, 2023.
- [7] Yuchen Cui, Siddharth Karamcheti, and et. al. “No, to the Right”—Online language corrections for robotic manipulation via shared autonomy. *arXiv preprint arXiv:2301.02555*, 2023.
- [8] Mariella Dimiccoli, Shubhan Patni, Matej Hoffmann, and Francesc Moreno-Noguer. Recognizing object surface material from impact sounds for robot manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022.
- [9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.
- [10] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.
- [11] Matthias Kerzel, Philipp Allgeuer, Erik Strahl, Nicolas Frick, Jan-Gerrit Habekost, Manfred Eppe, and Stefan Wermter. NICOL: A Neuro-inspired Collaborative Semi-humanoid Robot that Bridges Social Interaction and Reliable Manipulation, June 2023.
- [12] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *J. Mach. Learn. Res.*, 22:30:1–30:82, 2021.
- [13] Seung Hyun Lee and et. al. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3377–3386, 2022.
- [14] Mengdi Li, Xufeng Zhao, Jae Hee Lee, Cornelius Weber, and Stefan Wermter. Internally Rewarded Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, July 2023.
- [15] Corey Lynch, Ayzaan Wahid, and et. al. Interactive language: Talking to robots in real time. *arXiv preprint arXiv:2210.06407*, 2022.
- [16] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: A survey. *arXiv preprint arXiv:2302.07842*, 2023.
- [17] Long Ouyang, Jeff Wu, and et. al. Training language models to follow instructions with human feedback, March 2022.
- [18] Siddharth Patki, Andrea F. Daniele, Matthew R. Walter, and Thomas M. Howard. Inferring Compact Representations for Efficient Natural Language Understanding of Robot Instructions, March 2019.
- [19] Ian C Rankin, Seth McCammon, and Geoffrey A Hollinger. Robotic information gathering using semantic language instructions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4882–4888. IEEE, 2021.
- [20] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *Conference on Robot Learning (CoRL)*, volume 164, pages 894–906, 2022.
- [21] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2022.
- [22] Avi Singh, Albert Yu, Jonathan Yang, Jesse Zhang, Aviral Kumar, and Sergey Levine. COG: Connecting New Skills to Past Experience with Offline Reinforcement Learning. In *Conference on Robot Learning (CoRL)*, October 2020.
- [23] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, October 2022.
- [24] Jennifer Wakulicz, He Kong, and Salah Sukkarieh. Active information acquisition under arbitrary unknown disturbances. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 8429–8435. IEEE, 2021.
- [25] Hu Wang, Jianpeng Zhang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Uncertainty-aware multi-modal learning via cross-modal random network prediction. In *Computer Vision—ECCV*, pages 200–217. Springer, 2022.
- [26] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.
- [27] Andy Zeng, Maria Attarian, brian ichter, and et. al. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations (ICLR)*, 2023.
- [28] Xufeng Zhao, Cornelius Weber, and et. al. Impact Makes a Sound and Sound Makes an Impact: Sound Guides Representations and Explorations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2512–2518. IEEE, 2022.
- [29] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, April 2023.
- [30] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18:351–376, 2022.