2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10801833

Vision-Language Model-based Physical Reasoning for Robot Liquid Perception

Wenqiang Lai[1,2], Tianwei Zhang[1,2], Tin Lun Lam[1,2], Yuan Gao[1,2,†]

***Abstract*— There is a growing interest in applying large language models (LLMs) in robotic tasks, due to their remarkable reasoning ability and extensive knowledge learned from vast training corpora. Grounding LLMs in the physical world remains an open challenge as they can only process textual input. Recent advancements in large vision-language models (LVLMs) have enabled a more comprehensive understanding of the physical world by incorporating visual input, which provides richer contextual information than language alone. In this work, we proposed a novel paradigm that leveraged GPT-4V(ision), the state-of-the-art LVLM by OpenAI, to enable embodied agents to perceive liquid objects via image-based environmental feedback. Specifically, we exploited the physical understanding of GPT-4V to interpret the visual representation (*e.g.*, time-series plot) of non-visual feedback (*e.g.*, F/T sensor data), indirectly enabling multimodal perception beyond vision and language using images as proxies. We evaluated our method using 10 common household liquids with containers of various geometry and material. Without any training or fine-tuning, we demonstrated that our method can enable the robot to indirectly perceive the physical response of liquids and estimate their viscosity. We also showed that by jointly reasoning over the visual and physical attributes learned through interactions, our method could recognize liquid objects in the absence of strong visual cues (*e.g.*, container labels with legible text or symbols), increasing the accuracy from 69.0%—achieved by the best-performing vision-only variant—to 86.0%.**

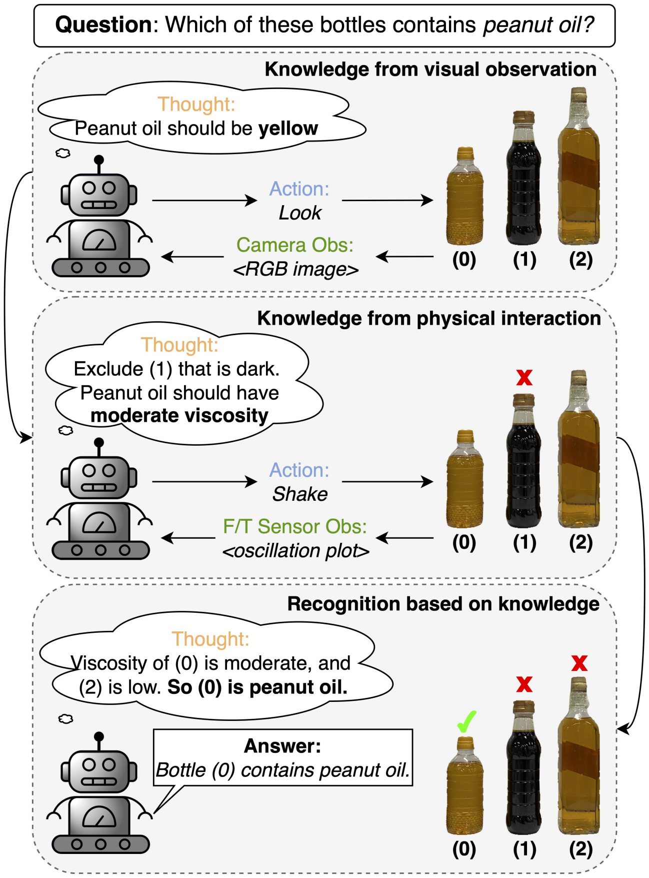

Question: Which of these bottles contains *peanut oil?*
Knowledge from visual observation
Thought:
Peanut oil should be **yellow**
Action:
Look
Camera Obs:
<RGB image>
(0)
(1)
(2)
Knowledge from physical interaction
Thought:
Exclude (1) that is dark.
Peanut oil should have
moderate viscosity
Action:
Shake
F/T Sensor Obs:
<oscillation plot>
(0)
(1)
(2)
Recognition based on knowledge
Thought:
Viscosity of (0) is moderate, and
(2) is low. **So (0) is peanut oil.**
Answer:
Bottle (0) contains peanut oil.
(0)
(1)
(2)

Fig. 1. Our method firstly predicts the visual and physical properties of the queried liquid object using commonsense knowledge (bolded text), and then explores liquids via *Look* and *Shake* actions interactively to estimate their properties from image-based feedback. Finally, the liquid with most consistent properties is selected as the answer.

I. INTRODUCTION

How would a human respond to a query like "*Bring me the milk please*"? Intuitively, humans would perceive visually the environment to look for the queried object. In case the object cannot be distinguished from visual information alone, we perform additional observations and reasoning over information from other modalities. For an intelligent robot to achieve human-like reasoning, the understanding of the feedback from the interactions in the physical world is essential [1]. In recent years, large language models (LLMs) have shown remarkable performance on commonsense and physical reasoning tasks [2], [3], making them potential reasoning models for robotic tasks, such as planning and manipulation [4]. To ground LLMs, prior works mainly relied on external modules to convert the multimodal environmental feedback into text. For example, [5] exploited large-language models (LLMs) to recognize the material of objects in a simulated environment by reasoning over textualized multimodal interaction feedback. However, the lack of direct perception of the environment limited the comprehension of the world. More recently, large vision-language models (LVLMs), such as GPT-4 V(ision) by OpenAI [6], enabled vision as an extra input modality in addition to language, which facilitated the incorporation of richer semantic knowledge. This opens up an interesting question: can LVLMs be used to reason over multimodal feedback from interactions in the physical world for robotic tasks? In this work, we used GPT-4V as our LVLM backbone model to explore a paradigm that adds image-based interaction feedback into reasoning the process of the LVLM to ground it within the physical world. Inspired

---

[1] Shenzhen Institute of Artificial Intelligence and Robotics for Society.
[2] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen.
[†] Corresponding author is Yuan Gao: gaoyuan@cuhk.edu.cn

This work is supported by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515110787 and 2024A1515012065), the Shenzhen Science and Technology Program (Grant No. JSGGKQTD20221101115656029), and Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects (Grant No. LGKCSDPT2024002.)

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

by [7], we allow the robot to act on a target object, and record the feedback in the form of images. Then, we feed the description of the taken action and the image-based feedback to the LVLM for inference. In this work, we evaluated this action-observation-reasoning paradigm with liquid objects, which require physical understanding to correctly recognize them, because the visual attributes (*e.g.*, color and texture) of liquid objects are either sometimes unobservable due to the opacity of the containers, or insufficient to distinguish a liquid from visually similar ones. For example, to distinguish a target liquid object from visually similar counterparts, the robot would shake the bottles, and reason over the time-series plots of liquid oscillations using its physical understanding to provide a qualitative estimate of their viscosity and select the one with the most consistent properties (Fig. 1).

We conducted a systematic evaluation of the proposed method on 10 common household liquids of varying container appearance, shape, and material. We established two settings with these liquids to simulate the variability in the appearance of liquids in the real-world. In the first setting, the original packaging labels remained intact, while these were removed or rendered invisible in the second setting. We used a robot arm as the embodiment and a wrist-mounted F/T sensor and a RGB camera to provide feedback. The main contributions of this work are summarized as follows:

- We explored a new paradigm that leverages LVLM to perceive and reason over physical response from liquid objects via image-based haptic feedback for the qualitative estimation of liquid viscosity.
- We demonstrated that by integrating both visual and haptic feedback, our method increased the accuracy to 86.0% in recognizing 10 common household liquids, compared to the 69.0% accuracy achieved by the variant using only visual feedback..

II. RELATED WORK

A. Large Models for Physical Reasoning

The knowledge of the physical properties of an object is crucial in many robotic tasks. Prior works studied learning-based methods for the estimation of physical properties from visual [8],[9] and other modalities [10],[11] of interaction data. However, these methods are task-specific and difficult to scale, as they require a substantial amount of training data. Recently, there has been an increasing interest in leveraging the rich world knowledge encoded in LLMs for reasoning.

LLMs, which reverse-engineer the world through the massive quantity of training text, achieved remarkable performance on a range of physical reasoning tasks [12]. However, since LLMs can only process texts, previous works relied on external modules [4], [13], [14] to provide textual description of the feedback from the environment. For example, [5] proposed a framework for robot planning in a simulated environment, where the robot agent plans to act in the environment to gather textualized object-centric physical properties from multimodal perception modules. However, the information provided by perception modules was conveyed via language, which may not provide the necessary context to comprehend the world, leading to inaccurate reasoning.

More recently, LVLMs were used to directly reason over the visual feedback to estimate a range of object-centric physical concepts, such as mass and deformability [15],[16]. These methods work by relating visually observable attributes, such as object material or category, to the intrinsic physical properties. Liquid objects have highly variable visual attributes, making the estimation of their intrinsic properties (*e.g.*, viscosity) challenging. In this work, we exploit the physical reasoning ability of LVLM to perceive the viscosity of liquids by reasoning over image-based haptic interaction feedback.

B. Robot Liquid Perception

The variability of liquids, combined with the way containers can alter their perceived shape and texture, posed great challenges to general object detection methods based on vision. As such, prior works leveraged non-vision interaction-based approaches to perceive liquids [17], [18], [19], [20]. These methods perceive the response of liquids to external motions (*e.g.*, shake, grasp, tilt, etc.) using various sensors, such as accelerometer, F/T sensor, and tactile sensor, based on which data-driven or physical analysis-based models were used for the estimation of physical properties (*e.g.*, viscosity) and/or classification of liquids.

Although data-driven methods [17],[18] reported promising results in liquid classification tasks, they lack of scalability and generalization to new liquids. In contrast, [19] proposed a method based on fluid dynamic analysis to estimate the volume, mass, and viscosity of liquids. However, this method requires exact knowledge about the container's geometry, which is impractical in many scenarios. Later, [20] combined physical analysis with a data-driven model to estimate the viscosity of liquids. The method works by firstly analyzing the dynamic tactile signals recorded after perturbations, and then using the extracted information (*e.g.*, rate of damping and oscillation frequency) to train a regression model. However, this method requires re-training to generalize to novel liquid containers.

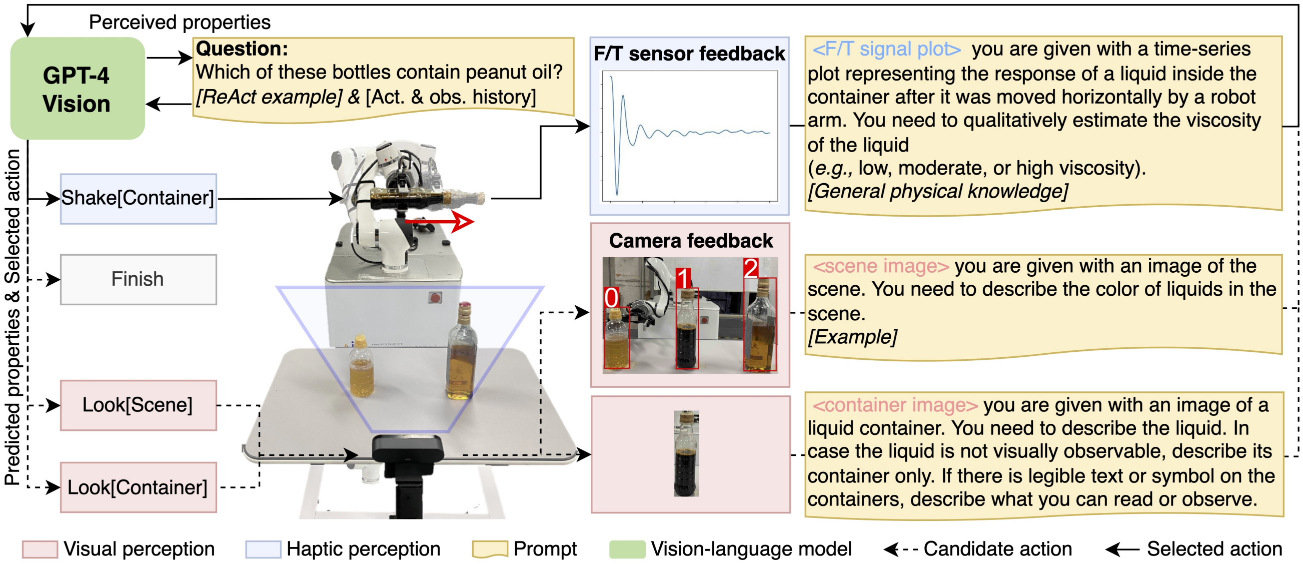
Despite prior works achieved remarkable results, they either suffer from the generalization and scalability issues, or difficulty in the real-world deployment. In contrast, we propose to leverage the commonsense and physical knowledge of LVLM to recognize liquid objects without any training or fine-tuning by reasoning over visual and haptic feedback.

III. METHODOLOGY

A. System Architecture

To enable interactive multimodal reasoning, we extended the reason-act (ReAct) framework proposed by [7] to allow the robot agent to act and perceive in the real world. Specifically for the liquid object perception and recognition tasks, we employed GPT-4V as an integrated perception, reasoning and planning model, and introduced an action space $A = \{\textit{Shake[Container], Look[Scene], Look[Container], Finish}\}$, where the first 3 actions are used to gather object-centric

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.



Question: Which of these bottles contain peanut oil? *[ReAct example]* & [Act. & obs. history]

<F/T signal plot> you are given with a time-series plot representing the response of a liquid inside the container after it was moved horizontally by a robot arm. You need to qualitatively estimate the viscosity of the liquid (*e.g.,* low, moderate, or high viscosity). *[General physical knowledge]*

<scene image> you are given with an image of the scene. You need to describe the color of liquids in the scene. *[Example]*

<container image> you are given with an image of a liquid container. You need to describe the liquid. In case the liquid is not visually observable, describe its container only. If there is legible text or symbol on the containers, describe what you can read or observe.

Fig. 2. To recognize a target liquid object, we prompt (yellow) GPT-4V to reason, act and perceive in a closed-loop. It first predicts the properties of the target liquid object using prior knowledge, and selects an appropriate action to perceive the actual properties from haptic (blue) and visual (pink) feedback. Haptic feedback is denoised and converted into time-series plot, while visual feedback is pre-processed to include object bounding boxes. Based on the predicted and perceived properties, GPT-4V plans for next action to gather more information. GPT-4V terminates the loop with the action *Finish* when the information is sufficient to recognize the target object. Dashed arrows represent potential action-reasoning paths not selected in the current loop.[1]

information, and the *Finish* action is used to terminate the action-reasoning loop and return the final answer. The actions are described in detail in Section III-B.

Our method works by mimicking human behavior in recognizing a target object. Given a set of objects, humans use their *internal world model* to predict the properties that the target object might possess, and then interactively explore the objects to gather multimodal feedback, from which the actual properties can be perceived. Then, the object with perceived properties that are consistent with the predicted properties is selected as the target object. Our method used LVLM as the *internal world model* to recognize liquid objects, leveraging its commonsense and physical understanding. Formally, let obj denote the target liquid object, q denote the prompting question (*e.g.*, "which of these bottles contains peanut oil?"), e denote the in-context example to guide the reasoning-action behaviour [7], and $c_t = (a_0, o_0, a_1, o_1, ..., a_{t-1}, o_{t-1})$ denote the context at time t, where o represents the textualized perceived properties after executing action $a \in A$. We prompt LVLM to first predict the properties of the queried object o'_t and then select an appropriate action a_t to perceive the corresponding properties o_t, which can be considered as a mapping:

$$F_{ReAct} : (q, e, c_t) \mapsto (o'_t, a_t) \quad (1)$$

after executing a_t, we prompt LVLM to perceive the object properties by reasoning over the image-based feedback $I_t \in \mathscr{T} = \{$*F/T signal plot*, *scene image*, *container image*$\}$, which can be regarded as visual question answering:

$$F_{perception} : (q_I, I_t) \mapsto o_t \quad (2)$$

where q_I is specific to the type of I_t (Fig. 2). The action-perception pair is then appended to the context in (1), $c_{t+1} = \{c_t, a_t, o_t\}$, to support the reasoning and acting in the next loop. A termination action *Finish* is available for the robot to finally return the answer when the perceived properties of a particular liquid object match the predicted properties of the queried liquid object $o_t \sim o'_t, \forall t$.

B. Multimodal Interactive Perception

To recognize liquid objects, complementing vision with haptic feedback from interaction is an intuitive approach for humans. Similarly, our method first observes the liquid objects using vision, and actively gathers additional physical knowledge about the objects via haptics.

1) Vision: Existing LVLMs, such as GPT-4V, tend to overlook fine details in high-resolution images and are prone to hallucinations when the scene is cluttered [6]. Therefore, we introduced actions *Look[Scene]* and *Look[Container]* to perceive coarse and fine-grained visual attributes, respectively. Color and shape are the most distinctive properties for object recognition [21], and since liquids are shapeless, we consider color as the coarse visual property to be observed through the action *Look[Scene]*. We pre-processed the scene images using the pre-trained open object detector OwlViT [22] to provide visual reference (*e.g.*, bounding boxes with index numbers on top) to the liquid objects, which has been shown to improve the visual question answering ability of LVLM [6]. To avoid hallucinations, we introduced the action *Look[Container]* that allows the robot to select the interested object in the scene and observe it more closely. For simplicity, we implemented *Look[Container]*

---

[1] Codes & prompts at https://github.com/laiwenq/VLM_liquid_perception

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

as a command that crops the interested object from the scene image, following [23]. Specifically, when the robot selects the action *Look[Container]*, a cropped image of the target container will be returned, from which the robot can observe fine-grained details (*e.g.*, transparency and legible text/symbols on the container) in the absence of distractors. To guide LVLM to generate the desired output format for *Look[Scene]*, we provided a hypothetical example, in which a generic label (*e.g.*, [Input Image]) replaces an actual image, to avoid information leakage during the evaluation. For the action *Look[Container]*, we prompt LVLM to simply provide a comprehensive description of the object in the cropped image without any in-context example.

2) Haptics: We used haptic feedback to capture the physical attributes, such as viscosity, of the liquid objects. Since our approach only focuses on reasoning over the haptic feedback from a high level, it is less sensitive to noises. Therefore, we employed a 6-DoF F/T sensor, rather than an expensive high-resolution tactile sensor used in [20], to collect haptic feedback during the interactions with the liquids. We designed the action *Shake[Container]* for the robot to shake the target container, and record the response of the liquid to the motion. Following [24], the robot uses its arm to move the target container, which is placed horizontally in the gripper, 10 cm along the axis connecting the bottom and the opening of the container. We employed a scripted motion since object manipulation is not the focus of this work. Since the motion was linear, and the force measurements were not sensitive enough to capture the motion patterns due to hardware limitations, we only recorded the torque measurements along the axis that is orthogonal to both the direction of motion and the gravitational force for 10 seconds while holding still the container. We applied a 5th-order low-pass filter with a cut-off frequency of 2 Hz to denoise the collected signal, and then normalized it to zero-mean and unit-variance following [18]. With the given context (*e.g.*, the action taken), the LVLM should interpret the image feedback (*e.g.*, plot of F/T sensor signal) using relevant physical understanding and provide an answer to the question. Humans require reference in reasoning about liquid viscosity. Likewise, we injected general physical knowledge, consisting of two descriptions of the expected oscillation patterns for low (*e.g.*, peaks with slowly decreasing amplitudes) and high viscosity (*e.g.*, peaks with rapidly decaying amplitudes) liquids to the prompt as references, following the experimental design in [25], in which the maxima and minima stimuli were presented to the observers before asking them to rate the viscosity of liquids.

IV. EXPERIMENTS

We used GPT-4V by OpenAI[6] as the LVLM backbone in our method. For more deterministic inference, we set the temperature parameter to 0. We conducted two experiments to answer the following questions: 1) Can GPT-4V correctly interpret the image-based haptic feedback using relevant physical understanding to estimate viscosity? 2) How does the available visual and haptic object-information affect the liquid recognition performance of GPT-4V? Our method was designed to work in an interactive manner in the physical world, however, we employed offline evaluation to save time for collecting haptic feedback in real-time to avoid system failures caused by irrelevant modules (*e.g.*, grasp failures).

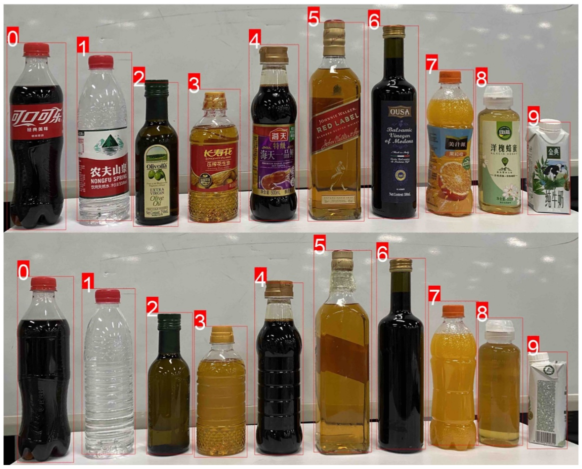


Fig. 3. Ten common household liquids evaluated in the experiments placed in a line on a table, each surrounded by a bounding boxed indexed between 0 and 9 from left to right: *coke, water, olive oil, peanut oil, soy sauce, whiskey, balsamic vinegar, orange juice, honey, milk*. (Top) Liquids in their original packaging with text and symbols on the labels. (Bottom) Same liquids with labels being removed or rendered invisible.

A. Hardware

We used a single-arm robot with a mobile base by Moying Technology and a wrist-mounted 6-DoF F/T sensor by Robitq with a sampling frequency of 100 Hz. A parallel gripper by Robotiq is mounted on the wrist as the end-effector. A Logitech Brio camera was placed on a tripod at a suitable pose, such that all liquid objects could be captured within the image. The liquids evaluated in this work were *coke, water, olive oil, peanut oil, soy sauce, whiskey, balsamic vinegar, orange juice, honey, milk*, which are commonly found in domestic environments with viscosity spanning from low (*e.g.*, *water*) to high (*e.g.*, *honey*). We measured the ground truth viscosity of the liquids using the NDJ-1 rotational viscometer. In contrast to most prior works that assumed containers of various liquids have the same shape and material, liquids evaluated in our experiments were contained in their original containers made of various materials (*e.g.*, plastic, glass, and paper) with various shapes and length. Relaxing this assumption is the key to the deployment in the real world, as in most cases the liquids have containers of variable geometry and appearance.

B. Data Collection

We collected images of the scene with different settings and the haptic feedback of each liquid separately, and provided these directly to GPT-4V when the corresponding action was chosen. As shown in the image of the scenes (Fig. 3), we established two distinct experimental settings to simulate the real-world variability of visual attributes in liquids and their containers, where one lacks control over the

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

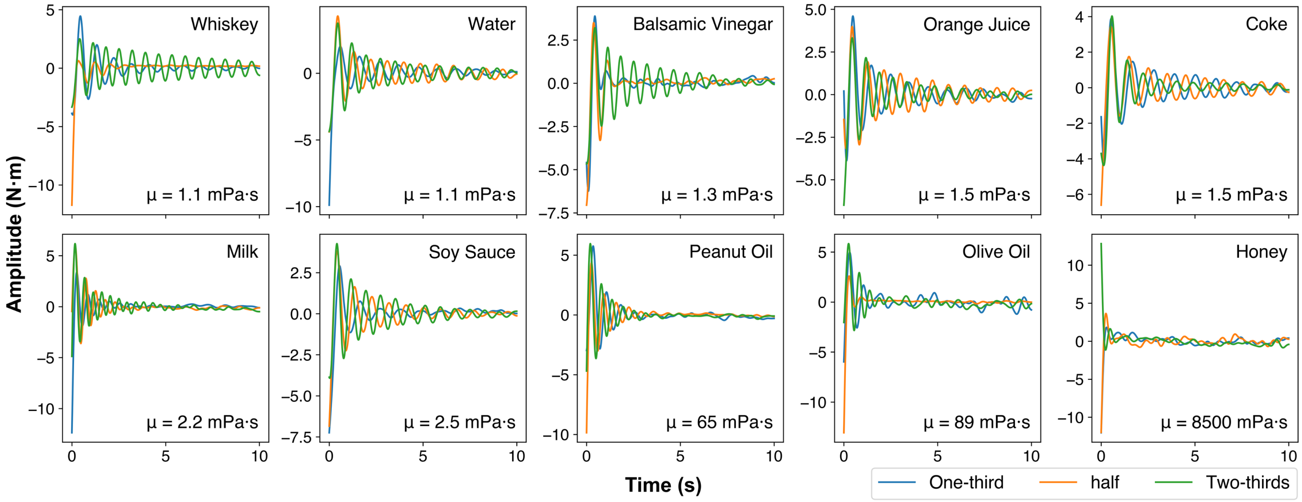


Fig. 4. The plots of 10-second F/T sensor signals collected after shaking the containers of 10 common household liquids filled at *one third*, *half*, and *two thirds* of capacity of their containers, sorted by their viscosity. Signals were filtered using a 5th order low-pass with a cut-off of 2 Hz and standardized.

presentation of these objects. Since the haptic signal might be too weak for perception when the fill level of liquids is either too low or too high. Following [20], we collected one haptic feedback from each liquid object at 3 fill levels in the middle range, namely *one-third, half, and two-thirds*. Fig. 4 showed the plots of the collected haptic feedback data of each liquid at different fill levels, sorted in ascending order by their ground truth viscosity. The fill levels were measured against the capacity of the containers using a beaker. As we do not deal with object grasp pose and force estimation, we simply hard-coded a grasping pose and force for each liquid. Following [18], the scripted grasp poses were set to around the middle region of the containers. The pose and force were determined empirically to avoid slippage.

C. *Relative Viscosity Estimation via Haptics*

To answer the first question, we tested the contextual haptic feedback understanding of GPT-4V by prompting it to observe the haptic feedback of two liquids and estimate their viscosity relation. Slight modifications were made to the haptic perception prompt in Fig. 2, where the input image consisted of two plots concatenated horizontally, and a question (*e.g.*,*"which one is more viscous?*) was appended at the end. To investigate how the injected physical knowledge affects the estimation of viscosity, we created and evaluated a prompt without external knowledge, which we refer to as plain prompt. We conducted 10 trials for each pair of liquids at each fill level. The accuracy was reported in Table I. Note that we excluded invalid outputs (*e.g.*, where GPT-4V refused to answer) from the calculation of accuracy.

As expected, we observed that the injection of knowledge increased the accuracy compared to the plain prompt across all fill levels. GPT-4V achieved the highest accuracy of 79.9% with knowledge-enhanced prompt in estimating viscosity relation of two thirds-filled liquids, surpassing the result obtained using plain prompt by 13.5%. The difference between accuracy across different fill levels was insignificant with plain prompt, while a direct proportion between the accuracy and the fill level was observed when using knowledge-enhanced prompt. We argue that this relation could be attributed to the differences in the shapes of containers. When the bottles are two thirds-filled, the impact of the shape of bottles would be minimal as the height of liquid when placed horizontally would be higher than the bottleneck. When the bottles are filled to one-third or half, the liquid's sloshing motion will be affected by the shape of the bottleneck, which has a variable design across each liquid. From the error breakdown (Fig. 5), we observed that the accuracy increment

TABLE I

ACCURACY OF GPT-4V IN PREDICTING PAIR-WISE VISCOSITY RELATION OF LIQUIDS AT THREE FILL LEVELS WITH PLAIN AND KNOWLEDGE-ENHANCED PROMPTS.

	Accuracy (%)	
Fill level	Plain	Knowledge-Enhanced
One third	66.4	77.1
Half	67.8	77.5
Two thirds	66.4	79.9

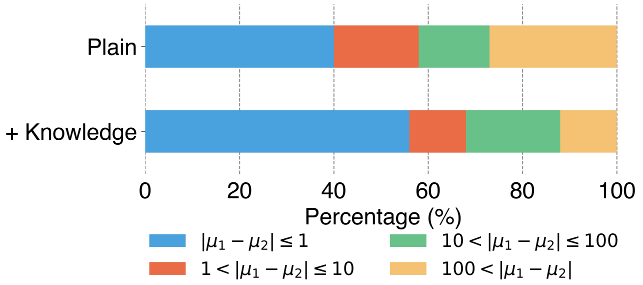


Fig. 5. The error breakdown of plain and knowledge-enhanced prompts the absolute ground truth viscosity difference of liquid pairs.

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

of the knowledge-enhanced prompt was mainly driven by the increased performance in estimating pairs of liquids with large differences in ground truth viscosity. The percentage error caused by liquid pairs with an absolute viscosity difference larger than 100 mPas was largely reduced.

D. Liquid Recognition via Multimodal Feedback

TABLE II

Recognition accuracy comparison between our method and its variants with partial action space.

Methods / Settings	Accuracy (%) W/o labels	Accuracy (%) W/ labels
Look[Scn.]	62.0	76.0
Look[Scn.]+Shake[Cnt.]	56.0	67.0
Look[Scn.]+Look[Cnt.]	69.0	**97.0**
Look[Scn.]+Look[Cnt.]+Shake[Cnt.](Ours)	**86.0**	93.0

We evaluated the performance of our method against its variants with partial action space (*e.g.*, some exploration action is unavailable) to understand the impact of object-centric information gathered from different exploration actions. In the liquid recognition experiments, we have two settings with different visually perceivable cues. As shown in Fig. 3, the first setting had the labels of all the containers facing the camera, so that the text and symbols on the labels can provide additional hints on the class of liquid. In the second setting (bottom), the labels of all the containers were removed to make sure the available visual cues were restricted to the container's color, shape, and material, as well as the color and texture of the content, in case of non-opaque containers. All liquids in the experiments were filled at around two-thirds of their respective capacity. We conducted 10 trials across all liquids in each setting.

As shown in Table II, our method achieved a cognition accuracy of 86.0% in the setting without labels on the containers, in which visual attributes were insufficient for correct recognition, outperforming all variants by large margins (17% to 30%), demonstrating the importance of physical knowledge. However, in both settings, we observed that the haptic perception combined with coarse visual perception (*Look[Scene]* with *Shake[Container]*) resulted in the lowest performance, even worse than using coarse visual perception alone. This might appear counter-intuitive at the first glance, however, since our method relied on coarse visual perception to select potential candidates for further interaction, misleading or insufficient information gathered in this first stage would strongly bias the future reasoning traces. This is consistent with the fact that the inclusion of fine-grained visual attributes largely increased the performance of all methods with action *Look[Container]*. Interestingly, in the setting with labels on the containers, we observed a small accuracy reduction of 4.0% in our method when compared to the vision-only variant with two visual perceptual actions, which we investigated in a case study later. The detailed breakdown of the predictions from our method and its variant without haptic feedback in the setting without labels on the containers was given in Fig. 6. The major limitation of the vision-only method was the confusion within liquids with similar colors. For example, *peanut oil* was confused with *whiskey*, and *honey*, while *soy sauce* was confused with *balsamic vinegar*. Whereas our method successfully allowed correct recognition of visually similar liquids with distinct viscosity (*e.g.*, *whiskey*, *peanut oil*, *honey*). However, no improvement was observed in the recognition of liquids with similar appearance and viscosity (*e.g.*, *soy sauce* and *balsamic vinegar*).

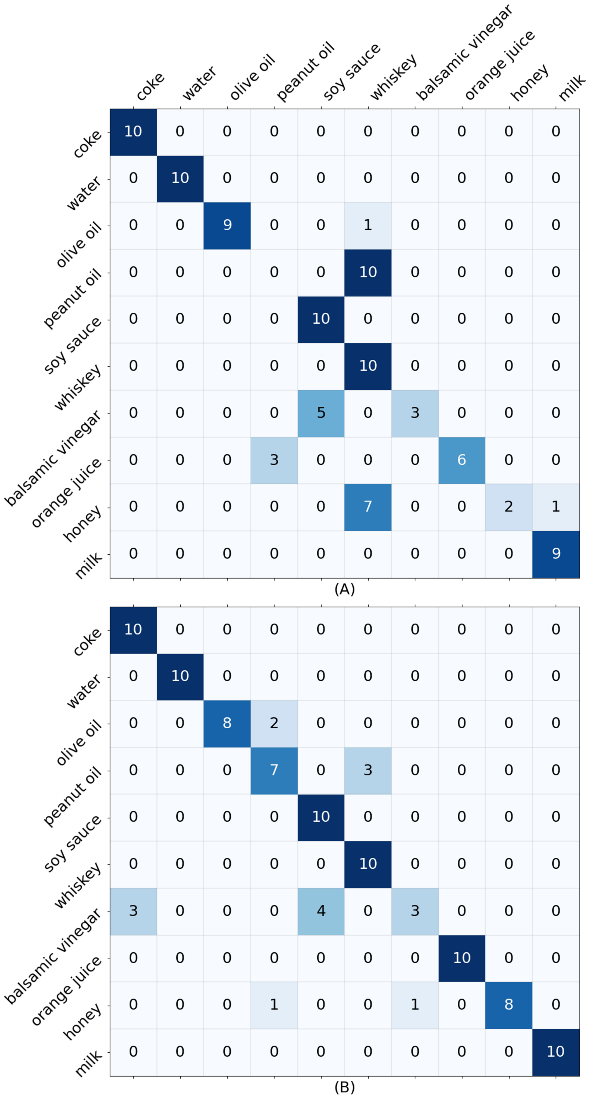


Fig. 6. The confusion matrices for the recognition results of 10 liquids from the method with only two visual perceptual actions (A) and our methods (B) in the setting without labels. 10 trials were conducted for each liquid, and invalid results (*e.g.*, where GPT-4V refused to answer) were excluded.

E. Case Studies

To understand how haptic feedback perception and physical understanding influenced the recognition of liquids across different settings, we provided the two case studies. Fig. 7 (top) showcased how the incorporation of haptic feedback combined with physical understanding contributed to the

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

[Setting: Containers without labels] - **Question:** Which of these bottles contains peanut oil?

Vision-Only	**Vision with Haptics**
Thought 1: Peanut oil should be pale yellow. I need to look at the bottles. Action 1: Look[Scene] Observation 1: [Scene image]. Thought 2: bottles (3) and (5) are potentially peanut oil based on their color. I need to take a closer look. Action 2: Look[Bottle 3] Observation 2: [cropped image of bottle 3] Thought 3: Now I will look at bottle (5). Action 3: Look[Bottle 5] Observation 3: [cropped image of bottle 5] Thought 4: Bottle (3) and (5) contain liquids with an amber color, which could potentially be peanut oil. Bottle (5) with handwritten label on the lid suggests a specialty or homemade product, which could align with peanut oil. Action 4: Finish[Bottle 5] ✗	*[Same reasoning traces as in Vision-Only]* Thought 3: Both bottle (3) and (5) have amber color matching the appearance of peanut oil. I need to shake both bottles to observe their viscosity and make a decision. Action 3: Shake[Bottle 3] Observation 3: [Oscillation plot of bottle (3)] Thought 4: Now I will shake bottle (5). Action 4: Shake[Bottle 5] Observation 4: [Oscillation plot of bottle (5)] Thought 5: Bottle (3) has a moderate to high viscosity, and bottle (5) has a low viscosity. Peanut oil typically has a moderate viscosity, so bottle (3) is more likely to be peanut oil. Action 5: Finish[Bottle 3] ✓

[Setting: Containers with labels] - **Question:** Which of these bottles contains honey?

Vision-Only	**Vision with Haptics**
Thought 1: Honey should have a golden color. I need to look at the bottles. Action 1: Look[Scene] Observation 1: [Scene image]. Thought 2: Bottles (3) and (5) are potential candi- dates as they have a golden yellow and amber color respectively. I need to take a closer look. [Collapsed reasoning traces for Look[Bottle 3] and Look[Bottle 5]] Thought 4: Bottle (3) contains peanut oil and bottle (5) contains whisky. I need to look at remaining candidates. [Collapsed reasoning traces for exhaustive search on remaining bottles] Thought 9: Bottle (8) contains a golden-yellow viscous liquid labeled as ” ACACIA HONEY,” which matches the typical attributes of honey. Action 9: Finish[8] ✓	[*Same reasoning traces as in Vision-Only*] Thought 4: Bottle (3) contains peanut oil and bottle (5) contains whiskey. I need to shake them to make a decision. [Collapsed reasoning traces for Shake[Bottle 3] and Shake[Bottle 5]] Thought 6: Bottle (3) has a golden-yellow color and moderate to high viscosity, so it is most likely to be honey. Bottle (5) has been ruled out as it contains whiskey and has a low viscosity response to shaking. Action 6: Finish[Bottle 3] ✗

Fig. 7. Reasoning traces of Vision Only (left) and Vision with Haptics (right) in recognizing *peanut oil* (top) in the setting without labels, and *honey* (bottom) in the setting with labels on the containers. Correct and misleading information were marked green and yellow, respectively.

separation of visually similar liquid objects in the setting without labels. In the reasoning process to recognize *peanut oil*, the vision-only method successfully identified bottle (3), which was the correct answer, as a potential candidate along with bottle (5), which contains *whiskey*, because these have an amber color that is typical for *peanut oil*. The vision-only system observed the handwritten label on the lid of bottle (5), which was assumed to be evidence of a homemade product. Without additional information, bottle (5) was wrongly returned as the answer. In contrast, with the incorporation of haptic feedback, the robot correctly recognized bottle (3) as *peanut oil*, which has a moderate to high viscosity that is more consistent with the knowledge of GPT-4V about *peanut oil* compared to bottle (5) that exhibited low viscosity. As seen previously, the inclusion of haptic feedback caused decrements in accuracy in the setting where there are clear symbols and legible texts on the containers. Fig. 7 (bottom) showcased the reasoning traces in recognizing *honey*, which mainly contributed to the accuracy decrement of our method in the setting with labels. We observed that both methods wrongly identified bottle (3) and (5) as potential candidates from the initial observation of the scene, failing to include the correct answer. Upon closer observations, bottle (3) and (5) were discovered to contain *peanut oil* and *whiskey*, respectively. The vision-only method then proceeded with an exhaustive visual search, and finally found bottle (8) was labelled ”Acaia Honey” and had a golden-yellow color that matched typical honey. Whereas our method proceeded to *Shake* the candidates, ignoring the observed inconsistencies, and finally recognized bottle (3) as *honey*, because of its moderate to high viscosity response. From this failed trial, we observed that our method is limited at distinguishing both visually and physically similar liquid objects, such as *peanut oil* and *honey*. The reason was that GPT-4V can only estimate the viscosity of liquids qualitatively, which was insufficient to describe the difference between *peanut oil* and *honey*.

V. DISCUSSION

The state-of-the-art LVLM, such as GPT-4V employed in this work, could potentially serve as the reasoning model for more generalized intelligent robot systems. As the reasoning

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.

traces revealed, GPT-4V can effectively reason over the image-based response of a liquid using its physical understanding to roughly estimate the viscosity, similar to humans whose perceptional sensing ability only allows qualitative description of the physical properties. Although this estimation is only qualitative, it is sufficient for humans to separate common liquids that are visually similar but differ largely in viscosity. As expected, GPT-4V exhibited similar reasoning traces in the liquid perception and recognition tasks as seen above. Although a direct comparison with prior works was impossible due to the variability in experimental settings, we would like to highlight that our method achieved promising zero-shot performance on liquid viscosity estimation and recognition tasks, comparable to the results obtained by task-specific models in the related works.

VI. CONCLUSION AND FUTURE WORKS

In this work, we demonstrated how LVLM can be used to reason about the physical properties of liquid objects, and recognize them interactively. We proposed a method allowing LVLM to act in the physical world to gather multimodal feedback, and evaluated it on a liquid recognition task, which is a challenging problem in robotics. The experimental results suggested that GPT-4V was capable to qualitatively estimate the viscosity of liquids from the their physical responses to external perturbation. Leveraging the knowledge about common liquid objects gained from pre-training, GPT-4V could recognize them by reasoning over the perceived visual and physical properties. A major limitation of this work is that GPT-4V is subject to continual updates that can affect model behavior, introducing a degree of unpredictability in the performance. Also, the impacts of lighting conditions and camera poses on the results, as well as the inference latency, were not studied as these are largely affected by the performance of the pre-trained backbone model, which is out of the scope of this work. Future works should consider evaluating a broader range of LVLMs to validate our approach's efficacy. In addition, the potential extension of our method to more modalities (*e.g.*, infer temperature from thermograms) would allow the robot to follow user requests more accurately, enhancing human-robot interactions.

REFERENCES

[1] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, p. e253, 2017.

[2] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.

[3] A. Talmor, O. Yoran, R. L. Bras, C. Bhagavatula, Y. Goldberg, Y. Choi, and J. Berant, "Commonsenseqa 2.0: Exposing the limits of ai through gamification," *Datasets and Benchmarks Track, Thirty-fifth Conference on Neural Information Processing Systems*, 2022.

[4] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *The Sixth Annual Conference on Robot Learning*, 2022.

[5] X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter, "Chat with the environment: Interactive multimodal perception using large language models," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023.

[6] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, 2023.

[7] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *The Eleventh International Conference on Learning Representations*, 2022.

[8] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," *Advances in neural information processing systems*, vol. 28, 2015.

[9] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik, "Learning visual predictive models of physics for playing billiards," *The Fourth International Conference on Learning Representations*, 2015.

[10] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2020, pp. 5633–5640.

[11] X. Guo, H.-J. Huang, and W. Yuan, "Estimating properties of solid particles inside container using touch sensing," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2023, pp. 8985–8992.

[12] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, and A. M. Dai, "Mind's eye: Grounded language model reasoning through simulation," *The Eleventh International Conference on Learning Representations*, 2022.

[13] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation*. IEEE, 2023, pp. 11 523–11 530.

[14] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *The Seventh Annual Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[15] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[16] L. Li, J. Xu, Q. Dong, C. Zheng, Q. Liu, L. Kong, and X. Sun, "Can language models understand physical concepts?" *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[17] S. Chitta, J. Sturm, M. Piccoli, and W. Burgard, "Tactile sensing for mobile manipulation," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 558–568, 2011.

[18] P. Güler, Y. Bekiroglu, X. Gratal, K. Pauwels, and D. Kragic, "What's in the container? classifying object contents from vision and touch," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3961–3968.

[19] C. Matl, R. Matthew, and R. Bajcsy, "Haptic perception of liquids enclosed in containers," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2019, pp. 7142–7149.

[20] H.-J. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," *Robotics: Science and Systems XVIII*, 2022.

[21] S. Buetti, J. Xu, and A. Lleras, "Predicting how color and shape combine in the human visual system to direct attention," *Scientific reports*, vol. 9, no. 1, p. 20258, 2019.

[22] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 728–755.

[23] P. Wu and S. Xie, "V*: Guided visual search as a core mechanism in multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 084–13 094.

[24] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9118–9147.

[25] J. J. R. Van Assen, P. Barla, and R. W. Fleming, "Visual features in the perception of liquids," *Current biology*, vol. 28, no. 3, pp. 452–458, 2018.

Authorized licensed use limited to: Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on March 07,2026 at 12:24:19 UTC from IEEE Xplore. Restrictions apply.