2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Lightweight Visual Reasoning for Socially-Aware Robots

Alessio Galatolo[1*], Ronald Cumbal[1*], Alexandros Rouchitsas[1], Katie Winkle[1],
Didem Gürdür Broo[1], and Ginevra Castellano[1]

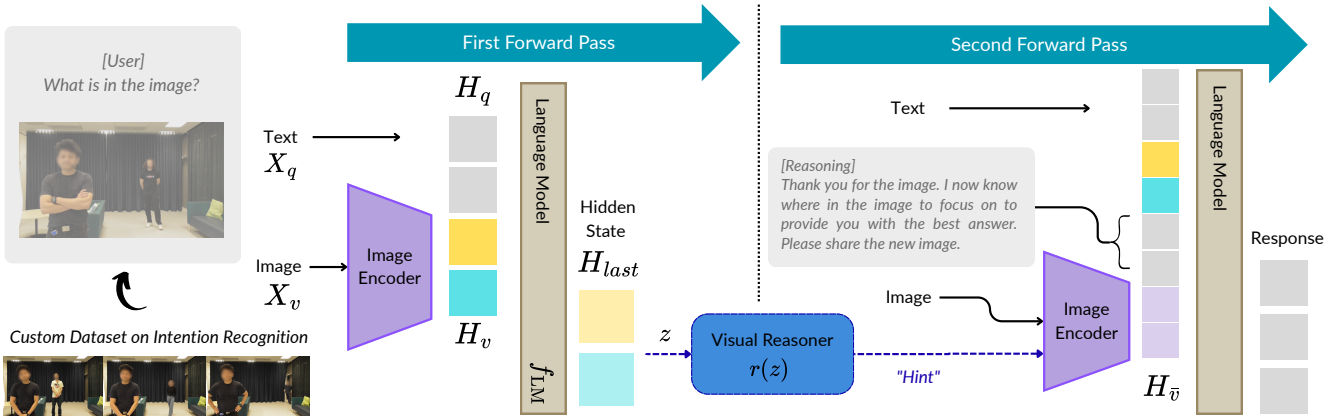

Fig. 1: **Overview of the Visual Reasoning approach**: A module connects an LLM's hidden states for image tokens back to the vision encoder through a gated MLP, creating a reasoning loop between text and vision. Training uses a two-pass strategy: the first pass extracts reasoning features from the LLM, and the second integrates them into the image encoding, reinterpreting the visual content in light of textual context and reasoning. Detailed description shown in Algorithm 1. A sample of the custom dataset on intention recognition during human-robot interactions is shown in the bottom left.

***Abstract*— Robots operating in shared human environments must not only navigate, interact, and detect their surroundings, they must also interpret and respond to dynamic, and often unpredictable, human behaviours. Although recent advances have shown promise in enhancing robotic perception and instruction-following using Vision-Language Models (VLMs), they remain limited in addressing the complexities of multimodal human-robot interactions (HRI). Motivated by this challenge, we introduce a lightweight language-to-vision feedback module that closes the loop between an LLM and the vision encoder in VLMs. The module projects image-token hidden states through a gated Multi-Layer Perceptron (MLP) back into the encoder input, prompting a second pass that reinterprets the scene under text context. We evaluate this approach on three robotics-centred tasks: navigation in a simulated environment (Habitat), sequential scene description (Mementos-Robotics), and human-intention recognition (our HRI dataset). Results show that our method improves Qwen 2.5 (7B) by** 3.3% **(less distance),** $+0.057$ **description score, and** $+2.93\%$ **accuracy, with less than** 3% **extra parameters; Gemma 3 (4B) and LLaVA OV 1.5 (4B) show mixed navigation results but gains** $+0.111, +0.055$ **and** $+10.81\%, +4.79\%$ **on the latter two tasks.**

I. Introduction

Integrating robots into human-shared environments goes beyond completing tasks and ensuring physical safety. It also demands a deep understanding, and often anticipation, of dynamic contexts. However, reasoning about and responding to such environments is inherently challenging, particularly when human behaviour is involved. For instance, robots operating in urban spaces must not only perceive their surroundings to achieve socially aware navigation but also employ effective communication strategies [1]. They are expected to react to spontaneous human behaviours [2] and make context-dependent decisions about how to engage with nearby individuals [3], [4]. These challenges—navigating diverse environments, understanding their characteristics, and managing human interactions—highlight the critical role of accurate environment reasoning in the successful deployment of robots in human-shared spaces.

Recent advancements have shown that Large Language Models (LLMs) can enable robots to better interpret and follow human instructions [5]. Similarly, Vision-Language Models (VLMs) have been proposed as a way to connect advanced visual perception capabilities to text, allowing robots to receive natural language instructions and being able to reason about their surroundings. Despite these developments, much of the existing research addresses these capabilities in isolation, overlooking the holistic integration required for robots to operate effectively in human-shared environments. In particular, the role of Human-Robot Interaction (HRI) is often overlooked [6]. To bridge this gap, our work examines how VLMs can be leveraged to strengthen robots' capacity to navigate diverse scenarios, with special attention to the challenges posed by complex human behaviours.

In particular, we assume that understanding human behaviour requires quite advanced reasoning and understanding

*Shared-first authorship
[1]Department of Information Technology, Uppsala University, Sweden. Mail correspondence to: alessio.galatolo@it.uu.se

of small cues [7]. We thus look at approaches that integrate reflection and reasoning mechanisms into VLMs. While techniques such as Chain-of-Thought (CoT) prompting [8] and structured reasoning [9] have significantly improved the performance of LLMs, their application in the context of multimodal models remains underexplored. Existing approaches to multimodal reasoning often rely on shallow combinations of visual and textual inputs, using visual information merely as context for textual reasoning rather than achieving deep integration of the two modalities [10]–[12]. To address this limitation, we introduce a novel reasoning module that establishes a direct feedback loop between visual and textual modalities. This deeper integration enables more robust interpretation of complex environments.

Our contributions are twofold:

- We propose a lightweight visual reasoning module that enables the language model to modulate the vision encoder, closing the loop between perception and interpretation—an underexplored architectural principle in current VLMs for robotics.
- We show empirical gains on scene description and intention recognition, and modest gains on navigation, analysed via ablations on image reuse, MLP removal, and order of input modality.

II. Related Works

LLMs have become powerful tools in the field of robotics, offering the ability to interpret complex instructions, reason through tasks, and communicate more effectively with humans using natural language [5], [13], [14]. At the same time, the integration of multimodal inputs—especially visual data—has enhanced the capabilities of LLMs beyond text-based understanding [15]–[18].

a) Vision models in robotic systems: Pre-trained VLMs, such as CLIP [19] and InstructBLIP [20], have played a pivotal role in enabling robots to process visual inputs for tasks such as object recognition and scene understanding [5]. For example, Kwon et al. [21] proposed combining LLMs with VLMs to facilitate grounded commonsense reasoning, allowing robots to actively perceive and interpret their environment. Similarly, Sermanet et al. [22] introduced RoboVQA, which leverages video input to support decision-making and visual understanding in complex, real-world scenarios. Furthermore advancing this line of work, Li et al. [23] introduced MMRo, a benchmark designed to evaluate robotic skills such as spatial reasoning, task planning, and safety awareness. Their findings indicate that even state-of-the-art models still face challenges in basic perceptual tasks, such as accurately identifying object attributes like colour, shape, material, and spatial location.

While these models have improved object and scene understanding in robotics, they largely overlook the interpretation of human behaviours, goals, or intentions —an ability considered fundamental to intelligent, cooperative systems [24], [25]. Research in HRI has attempted to address this challenge by studying human engagement [26], turn-taking [27], and interactions involving multiple participants [4]. However, the few efforts that apply vision models in these contexts show that even state-of-the-art systems continue to struggle in open-ended scenarios [3]. To address this gap, our work focuses on developing a reasoning module specifically designed to enhance the robot's ability to interpret to human behaviour in complex, real-world settings.

b) Visual reasoning: Early visual reasoning approaches used attention mechanisms to boost Visual Question Answering (VQA) performance [28] or trained models directly to enhance visual reasoning skills [29]. More recent methods incorporate prompting strategies to better guide models toward relevant visual features [30]. For example, DDCoT [31] breaks questions into sub-questions and uses external VQA models to generate rationales. Liu et al. [32] proposed a closed-loop framework that combines imagination and single-step reasoning, allowing models to iteratively refine their answers without further training.

Beyond reasoning strategies, some studies have enhanced understanding by explicitly manipulating visual inputs. For instance, Jiang et al. [33] and Lin et al. [34] used bounding boxes during inference or training to help models focus on relevant objects, improving reasoning performance. Shao et al. [35] extended this by jointly processing raw and box-annotated images to guide and strengthen the reasoning process. Other approaches, such as Image-of-Thought prompting [36], enable models to extract and generate both textual and visual rationales. Building on this, Zhang et al. [37] introduced a method for reasoning over multiple images by comparing visual similarities and differences.

Despite these contributions, we argue that, current approaches still rely heavily on textual generation for reasoning, with a shallow integration of visual information. Most visual manipulations, such as bounding boxes or segmentation, depend on external, pre-defined tools. As a result, the VLM itself lacks the ability to freely manipulate visual inputs, limiting its adaptability at inference time and preventing test-time scalability, as seen in text-only LLMs [9]. Moreover, these tool-based interventions break the end-to-end gradient flow, complicating the training and refinement of cross-modal reasoning capabilities. Instead, we propose a method that lets the LLM define manipulations independently by connecting its outputs directly to the vision encoder through a simple Multi-Layer Perceptron (MLP) module.

III. Methodology

We propose an approach to allow VLMs to do *cross-modal* reasoning by introducing a lightweight visual reasoning module that connects the language understanding component to the visual encoding process. We release our training and evaluation code at `https://github.com/alessioGalatolo/VLM-Reasoning-for-Robotics`.

A. Architecture choice

Starting with Flamingo [38], the prevailing approach for (open) VLMs has been to combine a pretrained vision encoder (e.g., CLIP [19]) with a pretrained or fine-tuned LLM

through alignment training [6]. Our method is specifically designed for such architectures and is compatible with widely used models, including LLaVA-OneVision [39], Qwen 2.5 VL [40] and Gemma 3 [41]. Moreover, due to its minimal requirements, our approach can potentially be extended to a broader range of architectures beyond these examples.

B. Visual reasoning module

This module is attached to the final layer of the language model, receives its hidden representation and connects it to the vision encoder, effectively forming a reasoning loop between the two architectures. More specifically, given an input sequence of a prompt (i.e., a *query*) and an image, we take the hidden states relative to the image, after it was processed by the language model alongside textual information. The visual reasoner then projects that back into the input space of the encoder, adding it to the image. By adding reasoning information before the image is fed into the encoder, we enable the model to reinterpret visual content in light of textual context and reasoning. At the end of this process, a second forward pass is done with the original prompt and image, plus the newly encoded one. Figure 1 illustrates in detail this process.

For the architecture of the visual reasoning module, we adopt a gated MLP:

$$\sigma(W_g x) \odot W_p\left(\text{Dropout}\left(W_2 \cdot \text{GELU}(W_1 x)\right)\right)$$

The MLP's input and output dimensions match the hidden dimension of the LLMs; while the MLP's hidden dimension is set to double of the input/output dimension.

The Visual reasoner also comprises of a 'patch unmerger', used to project back from the LLM representation space into the number of patches expected by the encoder.

C. Training strategy

We illustrate our training procedure in Algorithm 1.

Algorithm 1 Detailed training procedure

Require: Dataset $\mathscr{D}$, language model f_{LM}, visual reasoner $\mathbf{r}$, visual encoder f_{VE}

Inject LoRA adapters into f_{LM}
for image x_v, text $x_q \in \mathscr{D}$ **do**
  First forward pass (LoRA enabled if available):
    $H_v \leftarrow f_{\text{VE}}(x_v)$
    Compute hidden states $H \leftarrow f_{\text{LM}}(x_q, H_v)$
    Extract visual hint $z \leftarrow H_{\text{last}}$
    Set image reasoning: $\mathbf{r}(z)$
  Second forward pass (LoRA disabled always):
    $H_{\bar{v}} \leftarrow f_{\text{VE}}(x_v + \mathbf{r}(z))$
    Compute prediction $\hat{y} \leftarrow f_{\text{LM}}(x_q, H_v, H_{\bar{v}})$
    Compute loss $\mathscr{L}(\hat{y}, \text{labels})$
  Update parameters of $\mathbf{r}$ and LoRA if available
end for

During training, we perform two passes for each step. In the first, the model is given the user query and the input image and performs a standard forward pass. From here, we take the final hidden state of the LLM, for the tokens corresponding to the image. We then process these representations through the visual reasoning module and the 'unmerger'. The result is then added to the original image after embedding, but before it is encoded. This produces a new encoding of the image that takes into account the feedback from the LLM.

In the second pass, we feed the user query, the original image, and the new image to the model. The loss is then computed (only) from this pass and is used for backpropagation. The loss is thus back-propagated from the end of the language model, to the vision encoder, to the visual reasoner—updating its weights while keeping the LLM and vision encoder *frozen*.

In our experiments, we also test the integration of LoRA layers [44] in the language model. We only enable these layers for the first pass, to aid the model in providing feedback to the vision encoder, and keep it disabled otherwise. The total training parameters amount to less than 1.7% of the original model and less than 3% when also including LoRA.

D. Training data

For training, we use the Visual-CoT dataset [35], which provides image–question inputs paired with a reasoning output. We intentionally selected a non-specialized dataset. While training on a robotics-specific dataset would likely yield higher gains, it would also bias the evaluation against the base model (which was not trained on such data) and risk narrowing the applicability of VLMs. Using a general dataset instead preserves a key advantage of VLMs, i.e., their ability to address a wide range of problems beyond domain-specific settings. To reduce computational demand for training, we resize all images to 360p resolution.

E. Inference

Suppose input <image, question>, an initial forward pass is done (with LoRA enabled if available) to generate the visual reasoning hint z. The output of the visual reasoner $r(z)$ is used to change embedding of the vision encoder. The LLMs is then passed both the old and the new image and it generates its reply through standard practices (see Figure 1).

F. Optimization and infrastructure

Optimisation was performed using the AdamW optimiser, with the model checkpointed at regular intervals. Training was conducted on 4×A100 (40GB), completing each stage in a full epoch using half-precision (bf16). We sweep over 7 learning rates $\eta \in [1e-2, 1e-5]$, equally spaced. We also experiment with placing the image before and after the prompt (details in Section IV-B). At the end, we select the top 2 configurations, using the MME-CoT as validation dataset [45], and merge them via linear interpolation [46]. Training lasted between 1h30 and 2h45, depending on configuration and model size.

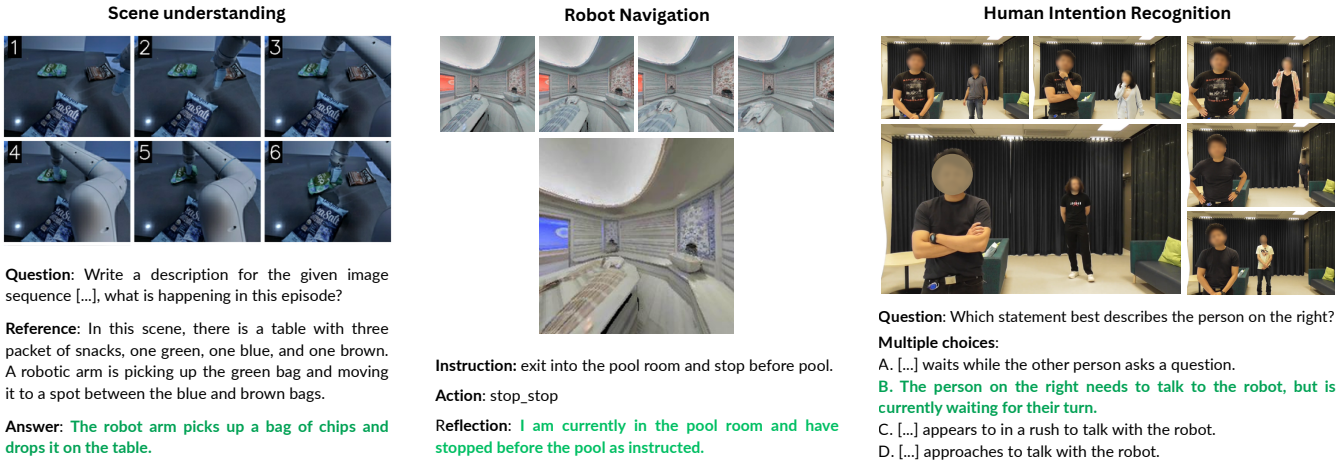

Fig. 2: Examples of the datasets use for evaluation. The Mementos-Robotics dataset [42] provides sequential images with scene descriptions; the Navigation benchmark [43] contains robot trajectories in simulation; and our human-intention recognition dataset captures interactions between humans and a social robot.

IV. Evaluation

We evaluate our approach against state-of-the-art VLM baselines and conduct several ablation studies. For the baselines, we consider three models: Qwen 2.5 VL 7B [40], Gemma 3 4B [41] and LLaVA OneVision 1.5 4B [39]. We then test the performance against three robotic-centred benchmarks: (i) robot navigation (ii) scene understanding (iii) human intention recognition. We use a greedy decoding strategy with no sampling for all the benchmarks.

- **Robot navigation**: We follow the pipeline from [47], where a robot navigates in the Habitat simulation [43] towards a goal specified in natural language. Robot actions are parsed from text (JSON) and can be: stay, move forward (0.25 step size), rotate left/right (15 degrees).
- **Scene understanding**: We use the Mementos-Robotics benchmark [42], which provides sequential images annotated with scene descriptions.
- **Human intention recognition**: We collected our own dataset, described in the following section.

A. Human intention recognition dataset

Since one of our main goals is to analyse human behaviour, with a particular focus on recognising human intentions, we require data with behaviour-specific annotations and a perspective resembling a robot's point of view. Existing robotics datasets often focus on pose prediction [48] or lack publicly available code/data [49]. Datasets such as UE-HRI [50] and JPL First-Person Interaction [51] meet some of these requirements, but they are limited in recording quality and behaviour variability. To overcome these limitations, we construct and annotate a new dataset[1].

Our dataset consists of audio-video recordings collected during a human-robot interaction study with 10 participants (3 female, 6 male, and 1 who preferred not to disclose gender). Participants had an average age of 30.9 years ($SD = 4.77$), most reported little prior experience with robots ($M = 2.00$, $SD = 1.33$ on a [1–5] scale), and all but one worked in engineering or technology-related fields.

The study featured the Furhat social robot, acting as a tourism assistant at an information desk. A *confederate*, playing the role of a tourist, initiated a conversation with the robot before participants entered the recording area. After signing a consent form, participants were then instructed to request information from the robot under three urgency conditions: *not rushed*, *somewhat rushed*, and *very rushed*. The duration of the ongoing interaction was shortened accordingly to naturally induce different levels of time pressure. Because conversational intervention can take multiple forms , the study was designed to observe whether and how participants chose to intervene depending on the urgency condition.

Our goal is to clearly specify participants' intentions while allowing them to express these intentions in a natural behaviour. This approach grounds their intentions and captures rich audiovisual data to study interaction strategies.

1) Annotation and benchmark adaptation: We annotated five types of participant behaviours from the video recordings: (1) waiting for their turn to speak with the robot, (2) approaching to interrupt the conversation, (3) calmly signalling intent to speak, (4) urgently signalling intent to speak, and (5) interacting with the robot while the confederate waited. In total, 188 events were annotated.

We pre-processed this set by converting each annotated caption into two multiple-choice questions, one referring to the inactive person (already engaged with the robot) and one to the intervening person (attempting to interact). First, we normalised captions by replacing explicit positional markers ('left'/'right' person) with a placeholder to build a pool of candidate templates. In a second pass, we restored the appropriate position for each instance and constructed four-option multiple-choice questions, ensuring that the correct

[1]Our application for ethical approval to the local review authority concluded that approval was not required, as the study does not fall under the provisions of the National Ethical Review Act.

description was always included alongside randomly sampled distractors from the option pool. Images were resized to a standard resolution (360p), and each question–answer pair was stored together with the corresponding processed image. The final dataset contains 376 samples, where each sample has a single associated frame and one question with four possible answers; the samples do not contain any textual information about the scene.

We do not make any splits of this dataset, as it is *only* used for the final evaluation.

B. Ablations and variations

To better understand the contributions of different components to the performance of our method, we conduct several ablation experiments. First, we remove the input image for the second embedding; in this scenario, the model receives the original image in the first pass and a completely artificial image (made by the LLM itself) in the second pass. Second, we remove the visual reasoning module, using only the hidden state from the LLM and project it back into the encoder's space (through the *unmerger*). Third, we experiment with ablating a stage, i.e., training only stage 1 or only stage 2. Finally, we test two variations in the order of inputs, passing the image either *before* or *after* the prompt (i.e., user query). Intuitively, passing the image *before* the prompt prevents its processed hidden states to take into account the prompt due to the causal masking used in LLMs.

C. Metrics

To provide a comprehensive evaluation of our method, we use a variety of metrics across benchmarks. For the navigation benchmark, we report the final distance to the goal, averaged across episodes. For the Mementos benchmark, which requires open-ended scene descriptions, we adopt an LLM-as-a-judge approach, scoring each generation from 1 to 5 based on its overlap with the ground-truth. For our intention recognition benchmark we opted for multi-choice questions and thus report accuracy as the evaluation metric.

V. Results

Table I reports the performance of our method compared to plain model baselines on all three tasks. On the Qwen 7B backbone, our approach achieves consistent improvements across all metrics: the final distance to goal in navigation is reduced (from 7.787 to **7.530**), open-ended description scores on Mementos increase (from 2.261 to **2.318**), and accuracy on intent recognition rises (from 34.04% to **36.97%**). For Gemma 4B, improvements are less uniform: our method provides a substantial boost in Mementos score (1.693 to *1.804*) and intention accuracy (20.84% to *31.65%*), but navigation distance slightly worsens (7.977 to 8.014). A similar pattern emerges for LLaVA 4B: our method improves Mementos score (from 2.201 to *2.256*) and intention accuracy (from 20.74% to *25.53%*), while navigation distance again slightly worsens (from 7.832 to 8.114).

Taken together, the results show that the proposed visual reasoning module yields the largest relative gains on scene description (Mementos) and human intention recognition, especially for models with lower initial performance.

	Navigation	***Mementos***	***Intentions***
Model	**Distance ↓**	**Score ↑**	**Accuracy ↑**
Plain Gemma 4B	*7.977*	1.693	20.84%
Ours (Gemma 4B)	8.014	*1.804*	*31.65%*
Plain LLaVA 4B	*7.832*	2.201	20.74%
Ours (LLaVA 4B)	8.114	*2.256*	*25.53%*
Plain Qwen 7B	7.787	2.261	34.04%
Ours (Qwen 7B)	**7.530**	**2.318**	**36.97%**

TABLE I: Performance of our method compared to the plain model. We highlight in **bold** the best performing model in each column and in *italic* the best version of each family.

	Navigation	***Mementos***	***Intentions***
Variation	**Distance ↓**	**Score ↑**	**Accuracy ↑**
Removing original	7.764	1.950	34.31%
No MLP	7.831	1.980	**37.50%**
Image first	**7.685**	**2.000**	28.46%
Prompt first	8.056	1.744	25.53%

TABLE II: Ablation studies using Qwen only. **Bold** highlights the best performing setting in each column.

The improvements are less significant (or sometimes negative) for the navigation task on Gemma and LLaVA. Here, we do a manual analysis of the generations, and for Gemma, we reduce this to a general difficulty of the model (both plain Gemma and our approach) in producing properly formatted output (i.e., JSON with action to take), resulting in the agent often skipping a 'step' due to the missing or incorrectly formulated action. On the case of LLaVA, we notice a very frequent refusal behaviour from the model (which does not happen in the other tasks). The Qwen backbone consistently benefits from our method, with gains in all three domains.

A. Ablation study

To isolate the contributions of each component, Table II presents results of several ablations using the Qwen backbone. Removing the original image in the second pass or ablating the MLP from the visual reasoner both lead to marked drops in performance on all benchmarks except intention recognition where the accuracy is still competitive at 34.31% without the original image and 37.50% without the MLP, compared to 36.97% with the full method. We attribute this effect to the 'patch unmerger' which is still a learned component (as it is strictly necessary for the method) who could partially compensate for the absence of the MLP. The navigation and Mementos metrics degrade in both ablations.

Ablating the order of input modalities (passing the image before or after the prompt) highlights how our initial assumptions were mistaken. The "image first" variant achieves 28.46% accuracy on intentions, while "prompt first" fares lower at 25.53%. Both are far below the complete method.

Model	Avg TFLOPs ↓	Samples/sec ↑	Peak memory (GB) ↓
Baseline	7.06	4.24	15.9
Ours	20.39	1.27	16.32

TABLE III: Resource consumption of our method compared to baseline. We report the average numbers when evaluating the Qwen model on our intentions dataset, tested on consumer hardware: NVIDIA RTX 3090.

The "prompt first" variant also scores lower on both the other two benchmarks.

In summary, the results confirm that both the use of the original image in the second pass and the presence of the MLP-based visual reasoner are necessary for best performance. The order in which the input modalities are fed into the LLM also matters, with "image before prompt" outperforming other settings.

B. Baselines

For our method, we experiment with placing the image before or after the user query, expecting increased performance for the latter (due to causal masking). Unexpectedly, our experiments reveal that our method performs best when the image appears before the user query. We attribute this issue to how the VLMs were initially trained, preferring one particular structure over the other. We empirically confirm this on the *base* models and report up to 30% performance degradation when swapping image and user query.

We further test another variation in the baselines. Here, we want to establish whether the improved performance of our method could be due to using two images instead of one. We thus provide our baselines two times the original image, in the same template as our method. This test yields even greater performance degradation, further supporting the effectiveness of our method.

Table I reports only the best results for the baselines.

C. Resource Consumption

Table III reports the computational overhead introduced by our method relative to the baseline, measured on an NVIDIA RTX 3090 during evaluation of the Qwen model on the human-intention recognition dataset. The additional cost is a direct consequence of the dual forward-pass design: performing two passes through both the vision encoder and the language model roughly triples the average TFLOPs (from 7.06 to 20.39) and reduces throughput from 4.24 to 1.27 samples per second. Despite this increase in compute, the memory footprint remains modest, rising by less than 3% (from 15.9 to 16.32 GB), which is consistent with the lightweight nature of the visual reasoning module and the fact that the additional parameters account for less than 3% of the original model. The increased latency is therefore primarily attributable to the extra inference pass rather than to any substantial growth in model size. Crucially, the method remains deployable on a single consumer-grade GPU, and a throughput of over one sample per second is sufficient for real-time deployment in robotics applications such as human-intention recognition and navigation, where perception typically operates at low frequencies. For scenarios demanding higher throughput, further optimisation through quantisation or hardware acceleration could be readily applied without altering the method itself.

VI. Discussion

Our results provide empirical evidence that introducing a lightweight visual reasoning module improves cross-modal reasoning across multiple robotics-centred tasks. This improvement is consistent across three different families of vision-language models, and particularly substantial for open-ended tasks like scene description and human intention recognition. In this section, we discuss both task-specific observations and broader implications of our design choices.

A. Interpretable cross-modal feedback loops

The core contribution of our method is the explicit feedback loop from the language model to the vision encoder. This loop, instantiated through a compact MLP and unmerging mechanism, enables the language model to modulate visual processing based on its understanding of the task and input prompt. Unlike most existing VLMs, where vision is passively embedded once and never updated, our approach reuses the language output to alter the visual embedding. Notably, our method does not require backpropagation through the vision encoder nor modifications to the base models, making it practical to integrate into existing VLM pipelines.

Further, our method provides a seamless integration of the two modalities, without breaking the gradient flow. In this initial work, we opted to use a general-purpose dataset (i.e., Visual-CoT [35]) to avoid any kind of cross-contamination and to show how our method is fit to improve a broader class of problems. However, it can be easily trained on more specialised data, potentially yielding far greater performance enhancement.

B. Performance across tasks

The three evaluated robotics-centred tasks differ in modality balance: navigation relies more heavily on structured outputs and spatial understanding; scene description requires contextual visual parsing; and intention recognition involves subtle social cues and multi-party reasoning. Our method yields consistent improvements in the latter two tasks, suggesting it is particularly effective when high-level visual semantics are critical.

For navigation, however, the gains are less consistent. In particular, performance with the Gemma and LLaVA backbone slightly declines, which we attribute not to failures in visual reasoning but to limitations in the model's ability to reply with well-structured output.

C. Design trade-offs and methodological lessons

Despite its simplicity, our method yields measurable gains with some trade-offs:

1) Image reuse in forward passes: We show that using the original image in both passes is essential—ablating this reduces performance across all tasks. This suggests that visual reinterpretation, rather than visual replacement, is a more stable strategy.

2) Ordering of modalities: Although our design places the image after the prompt to exploit causal masking, performance unexpectedly degrades for models initially trained with the reverse ordering.

3) Baselines using image duplication: We tested whether performance gains could be attributed to simply providing more visual input (e.g., duplicating the same image). These baselines performed worse, supporting the necessity of the modulated second image as opposed to brute-force repetition.

These findings indicate that introducing cross-modal feedback is not just a matter of adding more data or capacity, but of strategically closing the loop between perception and interpretation.

VII. Conclusion

We begin our work by investigating common challenges for robots situated in human-shared environments. Next, we target VLMs as general-purpose models, flexible in task definition and modality integration. Here, we introduce a novel approach for enhancing cross-modal reasoning in vision-language models through a lightweight visual reasoning module. Our intuition stands behind the idea that the 'default' encoding of the visual input is not aimed at grasping specific small cues that may be required by the context (e.g., predicting upcoming human behaviour), while the language model part is in principle capable of looking for such cues. Our method enables a feedback loop from language interpretation to visual perception by injecting dynamically generated reasoning hints into the vision encoder. Across three robotics tasks—navigation, scene understanding, and human intention recognition—our method consistently outperforms strong baselines with minimal additional parameters. More broadly, our results challenge the dominant feedforward paradigm in vision-language integration. By demonstrating that even frozen (encoder) models can benefit from guided visual reinterpretation, we provide a new tool for building more adaptive and context-aware robotic agents.

Our work highlights the value of architectural asymmetry and feedback in multimodal models—a principle well-known in embodied cognition but rarely realised in VLMs.

ACKNOWLEDGMENT

We thank Zixuan He for her work on implementing and debugging the LLaVA variant of our method.

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis, C3SE (Chalmers) partially funded by the Swedish Research Council through grant agreement no. 2022-06725. This research was supported by the Horizon Europe EIC project SymAware under the Grant Agreement No. 101070802.

References

[1] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "A survey on socially aware robot navigation: Taxonomy and future challenges," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1533–1572, 2024. [Online]. Available: https://doi.org/10.1177/02783649241230562

[2] D. Weinberg, H. Dwyer, S. E. Fox, and N. Martelaro, "Sharing the sidewalk: Observing delivery robot interactions with pedestrians during a pilot in pittsburgh, pa," *Multimodal Technologies and Interaction*, vol. 7, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2414-4088/7/5/53

[3] K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, and K. Ikeuchi, "Agreeing to interact in human-robot interaction using large language models and vision language models," *arXiv preprint arXiv:2503.15491*, 2025.

[4] M. Moujahid, H. Hastie, and O. Lemon, "Multi-party interaction with a robot receptionist," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022, pp. 927–931. [Online]. Available: https://doi.org/10.1109/HRI53351.2022.9889641

[5] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, "A survey on integration of large language models with intelligent robots," *Intelligent Service Robotics*, vol. 17, no. 5, pp. 1091–1107, 2024. [Online]. Available: https://doi.org/10.1007/s11370-024-00550-5

[6] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi, "Benchmark evaluations, applications, and challenges of large vision language models: A survey," *arXiv preprint arXiv:2501.02189*, vol. 1, 2025.

[7] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita, "Perceiving the person and their interactions with the others for social robotics–a review," *Pattern Recognition Letters*, vol. 118, pp. 3–13, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865518300771

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837.

[9] C. Snell, J. Lee, K. Xu, and A. Kumar, "Scaling llm test-time compute optimally can be more effective than scaling model parameters," 2024. [Online]. Available: https://arxiv.org/abs/2408.03314

[10] O. Thawakar, D. Dissanayake, K. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, M. Zumri, J. Lahoud, R. M. Anwer, H. Cholakkal, I. Laptev, M. Shah, F. S. Khan, and S. Khan, "Llamav-o1: Rethinking step-by-step visual reasoning in llms," 2025. [Online]. Available: https://arxiv.org/abs/2501.06186

[11] Z. Zhang, A. Zhang, M. Li, hai zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=y1pPWFVfvR

[12] R. Zhang, B. Zhang, Y. Li, H. Zhang, Z. Sun, Z. Gan, Y. Yang, R. Pang, and Y. Yang, "Improve vision language model chain-of-thought reasoning," 2024. [Online]. Available: https://arxiv.org/abs/2410.16198

[13] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[14] A. Zeng, B. Ichter, F. Xia, T. Xiao, V. Sindhwani, K. Bekris, K. Hauser, S. Herbert, and J. Yu, "Demonstrating large language models on robots." in *Robotics: Science and Systems*, 2023.

[15] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[17] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.10592

[18] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and

J. Tang, "Cogvlm: Visual expert for pretrained language models," 2024. [Online]. Available: https://arxiv.org/abs/2311.03079

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[20] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[21] M. Kwon, H. Hu, V. Myers, S. Karamcheti, A. Dragan, and D. Sadigh, "Toward grounded commonsense reasoning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5463–5470. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10611218

[22] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 645–652. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10610216

[23] J. Li, Y. Zhu, Z. Xu, J. Gu, M. Zhu, X. Liu, N. Liu, Y. Peng, F. Feng, and J. Tang, "Mmro: Are multimodal llms eligible as the brain for in-home robotics?" *arXiv preprint arXiv:2406.19693*, 2024.

[24] D. C. Dennett, *The intentional stance*. MIT press, 1989.

[25] A. Belardinelli, "Gaze-based intention estimation: principles, methodologies, and applications in hri," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 3, pp. 1–30, 2024. [Online]. Available: https://doi.org/10.1145/3656376

[26] N. Glas and C. Pelachaud, "Definitions of engagement in human-agent interaction," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 944–949. [Online]. Available: https://doi.org/10.1109/ACII.2015.7344688

[27] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021. [Online]. Available: https://doi.org/10.1016/j.csl.2020.101178

[28] Y. Zhou, T. Ren, C. Zhu, X. Sun, J. Liu, X. Ding, M. Xu, and R. Ji, "Trar: Routing the attention spans in transformer for visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2074–2084.

[29] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 61 501–61 513, 2023.

[30] Y. Zhang, S. Qian, B. Peng, S. Liu, and J. Jia, "Prompt highlighter: Interactive control for multi-modal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 215–13 224.

[31] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang, "Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023.

[32] J. Liu, Y. Li, B. Xiao, Y. Jian, Z. Qin, T. Shao, Y.-X. Ding, and K. Zhou, "Enhancing visual reasoning with autonomous imagination in multimodal large language models," *arXiv preprint arXiv:2411.18142*, 2024.

[33] S. Jiang, Y. Zhang, C. Zhou, Y. Jin, Y. Feng, J. Wu, and Z. Liu, "Joint visual and text prompting for improved object-centric perception with multimodal large language models," *arXiv preprint arXiv:2404.04514*, 2024.

[34] W. Lin, X. Wei, R. An, P. Gao, B. Zou, Y. Luo, S. Huang, S. Zhang, and H. Li, "Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want," *arXiv preprint arXiv:2403.20271*, 2024.

[35] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, "Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 8612–8642, 2024.

[36] Q. Zhou, R. Zhou, Z. Hu, P. Lu, S. Gao, and Y. Zhang, "Image-of-thought prompting for visual reasoning refinement in multimodal large language models," *arXiv preprint arXiv:2405.13872*, 2024.

[37] D. Zhang, J. Yang, H. Lyu, Z. Jin, Y. Yao, M. Chen, and J. Luo, "Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs," *arXiv preprint arXiv:2401.02582*, 2024.

[38] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 716–23 736.

[39] X. An, Y. Xie, K. Yang, W. Zhang, X. Zhao, Z. Cheng, Y. Wang, S. Xu, C. Chen, D. Zhu, C. Wu, H. Tan, C. Li, J. Yang, J. Yu, X. Wang, B. Qin, Y. Wang, Z. Yan, Z. Feng, Z. Liu, B. Li, and J. Deng, "Llava-onevision-1.5: Fully open framework for democratized multimodal training," 2025. [Online]. Available: https://arxiv.org/abs/2509.23661

[40] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[41] G. D. Gemma Team, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[42] X. Wang, Y. Zhou, X. Liu, H. Lu, Y. Xu, F. He, J. Yoon, T. Lu, G. Bertasius, M. Bansal, *et al.*, "Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences," *arXiv preprint arXiv:2401.10529*, 2024.

[43] X. Puig, E. Undersander, A. Szot, M. D. Cote, R. Partsey, J. Yang, R. Desai, A. W. Clegg, M. Hlavac, T. Min, T. Gervet, V. Vondrus, V.-P. Berges, J. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A co-habitat for humans, avatars and robots," 2023.

[44] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[45] D. Jiang, R. Zhang, Z. Guo, Y. Li, Y. Qi, X. Chen, L. Wang, J. Jin, C. Guo, S. Yan, B. Zhang, C. Fu, P. Gao, and H. Li, "MME-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=YZvefQVLJI

[46] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=6t0Kwf8-jrj

[47] Y. Duan and K. tang, "A navigation framework utilizing vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2506.10172

[48] K. Kedia, A. Bhardwaj, P. Dan, and S. Choudhury, "Interact: Transformer models for human intent prediction conditioned on robot actions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 621–628.

[49] V.-A. Le, B. Chalaki, V. Tadiparthi, H. N. Mahjoub, J. D'Sa, and E. Moradi-Pari, "Social navigation in crowded environments with model predictive control and deep learning-based human trajectory prediction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 4793–4799.

[50] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 464–472. [Online]. Available: https://doi.org/10.1145/3136755.3136814

[51] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2730–2737.