Look, Think, Understand: Multimodal Reasoning for Socially-Aware Robotics

Alessio Galatolo[1*] and Ronald Cumbal[1*]

***Abstract*— Robots operating in shared human environments must go beyond basic navigation and object recognition—they must also understand and respond to dynamic, socially embedded human interactions. While recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have shown promise in improving robotic perception and instruction-following, they often fall short in interpreting complex human behavior and intent. In this work, we explore how multimodal reasoning—specifically the integration of vision and language through reflective processes—can enhance robotic understanding in scenarios involving human interactions. We introduce a lightweight, bidirectional reasoning module that enables language-guided modulation of visual encoding, leading to a deeper interplay between linguistic and visual information. Here, after an initial forward pass with image and text, the model implicitly reasons on how the image could be manipulated to aid in giving a more accurate response. The image is then encoded once again while taking such considerations into account. We train our method on a diverse set of image-question datasets and evaluate it by constructing a dataset of real-world human interactions from the EGO4D egocentric video collection, which simulates a robot's perspective. Our experiments demonstrate that our method is able to bring consistent improvement over the plain VLM even when not trained on the task directly.**

I. INTRODUCTION

Integrating robots into human environments, particularly in shared spaces, requires more than just ensuring physical safety. It also involves fostering appropriate and socially acceptable behaviors. For robots to function effectively in these settings, they must not only interpret their surroundings but also understand how these spaces change over time. This goes beyond recognizing and analyzing physical objects; robots must anticipate and interpret how other agents, especially humans, influence both the environment and their own activities. For instance, delivery robots frequently encounter pedestrians who may either obstruct their path out of curiosity or assist them by providing guidance [1]. Similarly, service robots, such as those functioning as receptionists, must be able to determine whom to interact with, when to do so, and how to appropriately respond to different user intentions [2]. To handle these situations effectively, robots need to accurately recognize various human intentions and react in a suitable manner.

Recent advancements in robotics research have highlighted the role of large language models (LLMs) in enabling robots to interpret and execute human instructions [3]. A significant portion of research has further focused on improving robots' visual understanding to help them better interpret their physical surroundings. However, many of these studies tend to overlook the complexities of human-robot interaction, often concentrating primarily on social awareness in navigation [4]. When social interactions are addressed, they are typically limited to dyadic conversational scenarios with narrow objectives [5]. To address this gap, our research explores how a LLM with multimodal inputs e.g., Vision-Language Models (VLMs), can enhance robots' ability to interpret scenarios involving human interactions. While recent studies have begun exploring this direction, state-of-the-art methods still struggle with open-ended, real-world situations [6]. We pay particular attention to reflection and reasoning mechanisms present in these models. In fact, while Chain-of-Thought (CoT) prompting [7] and reasoning [8] have resulted in big increase in performance in LLMs, these techniques remain vastly underexplored in VLMs. The few works that claim *multimodal* reasoning, are actually just doing textual reasoning paired with vision input [9]–[11]. Motivated by this, we develop a novel multi-modal reasoning technique that is able to connect textual and visual modalities, with the possibility of extension to other forms of input as well. Here, the textual reasoning is able to effectively change how the image is processed in order to improve its embedded features and the resulting response.

To evaluate our approach, we analyze a dataset of egocentric video recordings that capture real-world interactions between the camera wearer and other people. These interactions simulate the kind of perceptual input a robot might receive through its camera system in everyday environments. While our broader goal is to improve reasoning capabilities across tasks, we specifically focus on the challenge of interpreting human intentions, a complex area that stands to benefit significantly from multimodal inputs. For example, recognizing internal human states often depends on a combination of verbal and nonverbal social cues, which must be interpreted together to understand intent accurately [6].

II. RELATED WORKS

Large Language Models 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 through natural language processing [3]. Simultaneously, the integration of multimodal inputs—especially visual data—has significantly enhanced the capabilities of LLMs in tasks involving visual understanding [12]–[14]. This convergence has paved the way for

*Shared-first authorship

[1]Department of Information Technology, Uppsala University, Sweden. Mail correspondence to: alessio.galatolo@it.uu.se or ronald.cumbal@it.uu.se

the development of Multimodal Large Language Models (MLLMs), which show promising potential for advancing robotics applications. In the sections that follow, we review recent developments at the intersection of these areas and examine how this synergy is shaping the future of robotic systems.

A. Reasoning and Multimodality with LLMs

Recent research has focused on improving the reasoning capabilities of LLMs through techniques such as Chain-of-Thought prompting and reflection-based methods. CoT prompting works by guiding models to break down complex problems into intermediate steps, leading to more accurate and interpretable outputs [7]. Reflection methods extend this idea by introducing feedback loops that allow models to iteratively refine their reasoning, identify errors, and produce more coherent responses. These strategies have been successfully applied in robotics [15] and in enhancing multimodal understanding [16], demonstrating their utility in complex, real-world scenarios.

Early efforts in visual reasoning with MLLMs employed attention mechanisms to improve Visual Question Answering (VQA) performance [17] or directly trained models to enhance visual reasoning abilities [18]. More recent approaches have incorporated prompting strategies to further focus MLLMs on relevant visual cues [19]. For instance, Zheng et al. [20] proposed DDCoT, a method that decomposes questions into sub-questions and uses external VQA models to generate rationales. Similarly, Liu et al. [21] introduced a closed-loop framework that combines imagination and single-step reasoning, enabling MLLMs to progressively refine their understanding and reach accurate conclusions without additional training or fine-tuning.

In addition to reasoning strategies, some studies have explored ways to enhance multimodal comprehension by explicitly manipulating visual inputs. Jiang et al. [22], for example, incorporated bounding boxes into the inference process to isolate target objects. Lin et al. [23] trained an MLLM to recognize images with overlaid bounding boxes, improving the model's reasoning capabilities. Similarly, Shao et al. [24] designed a system that simultaneously processes the original image and a version annotated with bounding boxes, using visual cues to guide and strengthen the reasoning process.

Despite these advances, most of these methods still rely primarily on textual generation, with only superficial integration of visual information. To address this limitation, Zhou et al. [25] proposed the Image-of-Thought (IoT) prompting framework, which allows MLLMs to autonomously extract and generate both textual and visual rationales, thereby enhancing multimodal reasoning. Further extending this idea, Zhang et al. [26] developed a method to guide models in answering complex questions that involve multiple image inputs by comparing similarities and differences across the visual content.

All these works go towards this idea of reasoning across modalities, however, none of them is able to achieve a free flow of information between text and vision. For example, drawing bounding boxes [22], [23] requires external tools to manipulate images, 'imagination' [21] requires a plethora of tools that need to be decided beforehand. The VLM is thus never able to 'freely' manipulate images and bring a similar effect of test-time scaling [8] that has emerged in text-only LLM. Further, such tool-based approaches break any gradient flow, making training and refining of cross-modal reasoning more challenging.

B. Multimodal Reasoning in Robotic Systems

The integration of LLMs with robotic systems has revealed transformative potential. Projects like PaLM-SayCan [27], [28] demonstrate how LLMs can be grounded in real-world robotic actions, enabling robots to interpret and execute natural language commands through physical behaviors. Further research has shown that incorporating multimodal inputs—particularly visual data—can significantly improve a robot's comprehension and generalization abilities.

Pre-trained VLMs, such as CLIP [29] and InstructBLIP [30], have played a pivotal role in enabling robots powered with LLMs to process visual inputs for tasks such as object recognition and scene understanding [3]. Much of this work has centered around Visual Question Answering (VQA) [31], where systems must respond to questions based on visual stimuli. For example, Kwon et al. [32] proposed combining LLMs with VLMs to facilitate grounded commonsense reasoning, allowing robots to actively perceive and interpret their environment. Similarly, Sermanet et al. [33] introduced RoboVQA, which leverages video input to support decision-making and visual understanding in complex, real-world scenarios. Li et al. [34] further contributed with MMRo, a benchmark designed to evaluate key robotic skills such as spatial reasoning, task planning, and safety awareness. Their findings indicate that even state-of-the-art MLLMs, such as Gemini-Pro, still face challenges in basic perceptual tasks, such as accurately identifying object attributes like color, shape, material, and spatial location. When it comes to interpreting human-robot interactions, recent studies have observed that while current models can manage clear, goal-directed interactions at the beginning of an exchange, they still face difficulties in open-ended scenarios [6].

While these studies demonstrate the advantages of combining multimodal inputs with LLMs for robotics, they also reveal important limitations. Most current efforts have focused on object recognition and scene interpretation, with less attention given to understanding the behaviors, goals, and intentions of humans or other agents in the environment. Even when human presence is considered, the emphasis is often on spatial grounding rather than on inferring intent or goal-directed behavior—an ability that is critical for effective human-robot interaction. Where intention understanding has been explored, recent findings continue to point to it as a significant and unresolved challenge.

III. Methodology

We propose an approach to enhance the multimodal reasoning capabilities of VLMs by introducing a lightweight visual reasoning module that allows the language understanding component to interface with the visual encoding process. Our objective is to propose a deeper, bidirectional interaction between vision and language components in generative settings that mirrors human processes. We have made our code publicly available.[1]

A. Architecture Choice

Contemporary VLMs often fall into one of two architectural paradigms: (i) alignment-based approaches that use pre-trained encoders like CLIP or BLIP to align vision and language in a shared embedding space (e.g., CLIP2, original BLIP), and (ii) integration-based approaches, such as LLaVA and Qwen-VL, where image features are injected into a language model to enable conditional generation, often using outputs from a vision encoder like BLIP-2.

Our method operates within the latter paradigm, which has emerged as the dominant strategy in recent high-performing models. Architectures like LLaMA-Adapter V2 [35], LLaVA [12], LLaVA-Next [36], and Qwen [37] bypass the need for an explicit intermediate representation like Q-Former by directly conditioning a decoder-only LLM on vision-derived tokens. This streamlines the integration of large-scale language modeling with multimodal reasoning, at the cost of reduced modularity and interpretability.

In contrast, models like BLIP-2 [30] and MiniGPT-v2 [38] introduce a Q-Former: a set of learnable query tokens that attend over frozen image encoders (e.g., ViT or CLIP) to extract vision-language aligned representations. While this architecture offers strong zero-shot performance and model reuse, it imposes a rigid, externally imposed structure on the interaction between modalities.

Our proposed mechanism introduces a soft, context-dependent interface between language and vision, which is missing in both CLIP2-style joint embeddings and Q-Former-based query extraction. Rather than querying image features via fixed learnable tokens, we allow the language model's internal representations—specifically, the last hidden state of the decoder—to dynamically modulate the vision encoder's class embedding via a reasoning module. This creates an organic feedback loop that does not break gradient flow, where linguistic interpretation guides visual re-encoding, which in turn shapes final generation.

Importantly, this mechanism is architecture-compatible with the LLaVA family of models (and easily extendable to similar models), which have become the *de facto* standard in open-source VLM development due to their simplicity, extensibility, and compatibility with decoder-only LLMs.

B. Visual Reasoning Module

We build upon the LLaVA-NeXT [36] architecture and introduce a learnable *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. The visual reasoner is conditioned on an initial joint image-text forward pass, and produces a reasoning vector or 'hint' that influences a new image embedding by modulating the class embedding that is prepended to the image. The class embedding acts as a pooling token and plays a critical role in the visual information summarization process. By adding reasoning information to this token, we enable the model to reinterpret visual content in light of linguistic context and reasoning. At the end of this process, a second forward is done with the new embedded image.

For the architecture of the visual reasoning module we test a gated multi-layer perceptron (MLP):

$$\mathbf{r}(\mathbf{x}) = \sigma(\mathbf{G}\mathbf{x}) \odot \mathbf{P}(f(\mathbf{x})) \tag{1}$$

where $f(\cdot)$ is a two-layer MLP with GELU activation and dropout, $\mathbf{G}$ and $\mathbf{P}$ are learned projection matrices, and σ denotes the sigmoid function. The resulting vector $\mathbf{r}$ is added to the frozen class embedding of the vision encoder before the second forward pass.

C. Training Strategy

We use two stages for training, shown in Algorithm 1. In the first stage, only the visual reasoner is trained and the VLM is kept frozen. In the second stage, we add a LoRA layer on top of the language model. This adapter is used to help the language model produce more meaningful visual reasoning hints and is *not* used in normal inference / generation.

During both stages, we adopt a two-pass training strategy. In the first pass, the model performs a standard vision-language reasoning step. This is done with the plain VLM in stage 1 and with LoRA enabled in stage 2. The final hidden state from the decoder (specifically, the last token representation) is passed to the visual reasoner. This produces a reasoning vector used to update the class embedding of the vision encoder.

In the second pass, with LoRA adapters *disabled* (i.e., we use the original VLM) in stage 2, the model re-encodes the image using the updated embedding and generates a response. Only the loss from this second pass 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. In the second stage, this is further propagated through the LoRA layer of the language model.

D. Inference

While we plan on extending and optimizing the inference procedure for generation, our current implementation closely resembles that of training.

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 the class embedding of the vision encoder. A second forward pass is done with the same input (original

[1] https://github.com/ronaldcumbal/uppsala_llm_hri

Algorithm 1 Two-Stage Training of Visual Reasoner and Language Model

Require: Dataset $\mathscr{D}$, number of epochs E

Initialize language model f_{LM}, visual reasoner $\mathbf{r}$

Stage 1: Pretrain visual reasoner r with frozen language model

for epoch $= 1$ to E **do**
  for batch $b \in \mathscr{D}$ **do**
    Compute hidden states $H \leftarrow f_{\text{LM}}(b)$
    Extract visual hint $z \leftarrow H_{\text{last}}$ ▷ last token hidden state
    Set image reasoning: $\mathbf{r}(z)$
    Compute prediction: $\hat{y} \leftarrow f_{\text{LM}}(b; \mathbf{r}(z))$
    Compute loss $\mathscr{L}(\hat{y}, \text{labels})$
    Update parameters of $\mathbf{r}$
  end for
end for

Stage 2: Joint finetuning of r and f_{LM} with LoRA adapters

Inject LoRA adapters into f_{LM}
for epoch $= 1$ to E **do**
  for batch $b \in \mathscr{D}$ **do**
    First forward pass (LoRA enabled):
      Compute hidden states $H \leftarrow f_{\text{LM}}(b)$
      Extract visual hint $z \leftarrow H_{\text{last}}$
      Set image reasoning: $\mathbf{r}(z)$
    Second forward pass (LoRA disabled):
      Temporarily disable LoRA adapters in f_{LM}
      Compute prediction $\hat{y} \leftarrow f_{\text{LM}}(b; \mathbf{r}(z))$
      Compute loss $\mathscr{L}(\hat{y}, \text{labels})$
      Re-enable LoRA adapters
    Update parameters of both $\mathbf{r}$ and f_{LM} (LoRA only)
  end for
end for

image and user query) with the addition of the new image. Thanks to the changed class embedding, a more informative encoding of the image is fed to the language model, thus producing a better output. An example generation is like the following:

> (User) <image> What is in the image?
> (Assistant) Let me begin by analyzing the image. Based on the question, the image would benefit from being manipulated like this: [manipulation] <new image> Based on the new image, the answer is: ...

Here, it is important to note how the **first** forward pass ends with "[manipulation]" and that "[manipulation]" is *not* a placeholder (i.e., the LLM will *not* explicitly describe the manipulation) but rather a dummy token passed to the LLM from which the reasoning hint will be extracted.

E. Dataset

Since our work focuses on interpreting and clarifying human interactions, we sought datasets that specifically include annotations related to such behavior. Additionally, we prioritized recordings captured "in the wild" that could resemble a robot's point of view. While the UE-HRI dataset [39] and the JPL First-Person Interaction dataset [40] met these criteria, they presented certain limitations: UE-HRI suffered from recording issues, and JPL contained too few interaction instances.

In contrast, the EGO4D dataset [41] offers a large collection of egocentric video recordings, densely annotated with narrated descriptions and various metadata. For our purposes, we leveraged the narration annotations, which include summaries and descriptions of events throughout the video. From this dataset, we extracted a subset of interactions initiated by individuals other than the camera wearer—highlighting how humans may interact with a robotic system. Examples of such interactions include: "Other person takes a basket from ego-person" or "Other person takes a card from ego-person." These were paired with broader video-level narrations such as "Ego-person was in a room, played a card game, and interacted with person A, B, and C." We excluded passive events (e.g., "Other person stands beside ego-person") to focus on more meaningful interactions. In total, we processed 460 annotated interaction instances.

However, the EGO4D dataset alone does not provide a sufficient number of examples to train our architecture effectively. To address this, we use the Visual CoT [24] dataset for training, reserving the EGO4D subset for evaluation only (also alleviating possible contamination between training/test). Visual CoT is a collection of different dataset, where samples are composed of natural language questions paired with images. We filter out corrupted images and ensure that each sample includes both visual and textual content. Finally, we balance (undersample) each dataset to have the same number of samples.

Each training instance is processed to form two types of prompts: a reasoning prompt (for the first pass) and an answer generation prompt (for the second pass). Input tokens corresponding to the reasoning prompt are masked in the label tensor to ensure that only answer tokens contribute to the loss.

F. Optimization and Infrastructure

During the initial training phase (Stage 1), only the parameters of the visual reasoner are updated. In a subsequent phase, we introduce LoRA adapters into the language model and jointly fine-tune both modules (Stage 2).

Optimization is performed using the AdamW optimizer with a learning rate of 1×10^{-5}. We checkpoint the model at regular intervals. The training was done on $2\times$A100 (80GB), where each stage completed a full epoch in half precision (bf16). Stage 1 lasted 12h40m, while stage 1 16h40m.

G. Evaluation

We use the subset of the EGO4D dataset for evaluation and compare all the variations of our method with the plain VLM. As the dataset does not contain any explicit question, we prompt the VLM with "*What are the human's intentions?*". We report the average loss and differences in perplexity between our method and the plain VLM baseline.

IV. Results

We show in Table I the performance of our method. Here, we observe how our method is able to improve on the performance of the plain VLM despite not being directly trained on the specific task/dataset. Here, we also see that Stage 2 brings better performance overall. However, this is at the expense of slower training and a higher parameter count.

One downside of our method is its inference speed, which is significantly slower than the plain VLM. This is a direct effect of the two forward passes, where in the second, our method also has to process two images, thus greatly increasing its context length. Nevertheless, this is in line with test-time scaling works where speed is often sacrificed in favor of better performance.

TABLE I

Results of our method compared to plain VLM. The perplexity difference is computed with respect to the plain VLM.

Method	Avg. loss ↓	Perplexity difference ↑	Time (s) ↓
Plain VLM	3.971	-	**84**
Ours - Stage 1 only	3.306	1.59	263
Ours - Stage 1 + 2	**3.148**	**1.77**	236

V. Conclusion and Discussion

In this work, we introduced a novel multimodal reasoning mechanism that enables deeper integration between vision and language for robotic systems. By establishing a dynamic feedback loop where linguistic context guides visual reinterpretation, our approach mirrors aspects of human processes. Through both architectural innovation and targeted evaluation on real-world interaction data, we demonstrate the potential of language-informed visual encoding for enhancing social awareness in robots. These findings not only highlight the value of bidirectional reasoning in multimodal models, but also open avenues for exploring richer forms of perception and interaction in human-robot collaboration.

Acknowledgment

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.

References

[1] 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, p. 53, 2023.

[2] Y. Xue, F. Wang, H. Tian, M. Zhao, J. Li, H. Pan, and Y. Dong, "Proactive interaction framework for intelligent social receptionist robots," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 3403–3409.

[3] 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.

[4] 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*, p. 02783649241230562, 2024.

[5] 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.

[6] 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.

[7] 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. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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.

[13] 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

[14] 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

[15] Y. Hao, F. Yang, and N. Fang, "Cora: A chain of robotic actions reasoning model for autonomous robotic arm manipulation," in *2025 11th International Conference on Automation, Robotics, and Applications (ICARA)*. IEEE, 2025, pp. 165–169.

[16] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[30] 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.

[31] R. Mirjalili, M. Krawez, S. Silenzi, Y. Blei, and W. Burgard, "Lan-grasp: Using large language models for semantic object grasping," *arXiv preprint arXiv:2310.05239*, 2023.

[32] 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.

[33] 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.

[34] 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.

[35] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[36] 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/

[37] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[38] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[39] 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

[40] 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.

[41] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 995–19 012.