Foundational Models for Robotics need to be made Bio-Inspired

Liming Chen[1] and Sao Mai Nguyen[2]

Abstract— **Foundation Models for Robotics (FMRs) promise to bring large-scale, generalist intelligence to embodied systems, yet they remain limited in their ability to integrate perception, action, and reasoning in physically grounded environments. In this paper, we argue that advancing FMRs requires drawing inspiration from biological systems—specifically human cognition, development, and sensorimotor learning. We outline five key bio-inspired principles for future FMRs: (1) memory architectures incorporating semantic, episodic, and procedural structures; (2) grounded structured reasoning, as exemplified by embodied chain-of-thought (E-CoT) processes; (3) integration of multimodal sensorimotor feedback, including touch and proprioception; (4) self-motivated learning through simulated play and intrinsic exploration; and (5) neural efficiency through sparse expert activation, functional specialization, and modular reasoning. These elements enable generalization, compositionality, and robustness—traits long demonstrated by humans but underrepresented in current robotic models. While this work does not address reliability and safety in depth, we identify them as essential future directions for developing trustworthy, human-aligned FMRs.**

I. INTRODUCTION

Foundation Models (FMs) [1] in Natural Language Processing (NLP) and Computer Vision (CV) refer to large-scale, pre-trained models that serve as a base for a wide range of downstream tasks. These models leverage massive datasets and self-supervised learning to develop general-purpose representations, which can then be fine-tuned for specific downstream applications.

In NLP, Large Language Models (LLMs), such as GPT series, PaLM, and LLaMA, have revolutionized the field by leveraging the transformer architecture [2], self-supervised pre-training paradigms and web-scale data. These models, often containing tens or hundreds of billions of parameters, demonstrate remarkable zero-shot and few-shot generalization capabilities across a diverse range of tasks [3], including dialogue systems, step-by-step reasoning, mathematical problem-solving, and code generation.

Although impressive, these foundation models lack sensorimotor skills, preventing them from directly perceiving, manipulating, or interacting with the physical world in an embodied manner. Such a contrast has been named **Moravec's paradox** [4] which explains why FMs, which can generate complex plans and solve abstract problems, struggle with basic sensorimotor tasks that human beings perform effortlessly. A toddler can walk, localize and recognize objects, and grasp items with almost no conscious effort.

[1]Liming Chen is with Liris, CNRS UMR 5205, Ecole Centrale de Lyon, 69134 Ecully, France and Institut Universitaire de France (IUF), France liming.chen@ec-lyon.fr

[2]Sao Mai Nguyen is with U2IS, ENSTA Paris, IP Paris, France nguyensmai@gmail.com

A natural and important question is whether these data-absorbent FMs can be extended to robotics and endowed with sensorimotor skills to interact with the real world. Such an extension could be a transformative leap in embodied AI, integrating high-level semantic understanding with low-level robotic control. Thanks to common sense knowledge in absorbing web-scale data, such an advancement could grant robots the ability to interpret and execute tasks through multimodal instructions. The extended FMs demonstrates human-like controllability via natural language interaction, perceptual reasoning to identify objects within the scene, and versatile, multi-task capability to execute a range of manipulative actions using a unified model. It exhibits strong generalization, adapting to diverse environments and object configurations, and incorporates long-horizon planning and skill composability, breaking down complex instructions into discrete, sequential actions—such as grasping, pouring, and handing over an object. By bridging the gap between visual understanding and embodied execution, this approach enhances robots' ability to act autonomously and flexibly in dynamic, real-world settings.

As such, an accelerating rapid move in AI and robotics is research on **Foundation Models for Robotics (FMRs)**, with increasingly intensive efforts to integrate sensorimotor capabilities and multimodal learning to enhance robotic autonomy and real-world generalization skills. The latest FMRs, such as **OpenVLA, π_0, and Octo**, represent a significant advancement in generalist robotic capabilities by integrating vision-language-action (VLA) models with large-scale real-world robotic data. **OpenVLA**[5], a 7B parameter model, is trained on 970,000 real robot demonstrations and outperforms larger models like **RT-2-X** (55B) [6], while maintaining adaptability across multiple robotic platforms. π_0, developed by Physical Intelligence, employs a flow-matching architecture built on vision-language models (VLMs) to enable smooth real-time 50Hz control across various tasks, including laundry folding and grocery bagging. **Octo**[7], a diffusion-based policy trained on 800,000 trajectories, provides flexible task execution by allowing natural language or goal image inputs, demonstrating high adaptability across nine robotic embodiments. These **FMRs** leverage scalable architectures, diverse real-world datasets, and hierarchical reasoning to push the boundaries of multi-task, generalist robotic intelligence, marking a shift toward more efficient, adaptable, and open-source robotic learning frameworks.

Despite recent progress, current FMRs still suffer from several limitations, including scarcity of robot data and its diversity, reliance on monolithic architectures, ineffi-

cient computation. In contrast, biological systems achieve remarkable adaptability with limited data, energy efficiency, and seamless sensorimotor integration. In this paper, we first discuss the limitations of current FMRs and explore how bio-inspired principles—such as structured memory, functional specialization, can enhance FMRs, making them more resilient, efficient, and autonomous, capable of skill acquisition and transfer.

II. LIMITATIONS OF CURRENT FMRS

FMRs are primarily built upon state-of-the-art Vision-Language Models (VLMs), which require vast amounts of robotic interaction data to learn representations that generalize across tasks [6]. However, unlike language models trained on web-scale datasets, robotic foundation models suffer from a severe data bottleneck, as acquiring diverse, high-quality demonstrations for embodied agents is expensive and time-consuming [8]. Furthermore, these models face high computational demands and brittle generalization, making their deployment in dynamic, unstructured environments particularly challenging. In addition, FMRs inherit critical shortcomings from VLMs, including hallucination—generating incorrect or unrealistic actions—and a lack of real-world grounding, which can lead to imprecise control and failure in physically interactive tasks [9].

III. ENDOWING FMRS WITH HUMAN LIKE MEMORIES

Human memory consists of several specialized subsystems that collectively support perception, learning, and behavior [10]. **Short-term memory** enables real-time reasoning and decision-making by holding transient information, while **long-term memory** stores accumulated knowledge and experiences. Within long-term memory, **semantic memory** encodes general knowledge (e.g., tool functions), **episodic memory** captures temporally situated personal experiences, and **procedural memory** supports motor skill learning and automatic behaviors (e.g., grasping or walking). **Meta-memory** governs self-awareness and regulation of memory processes, guiding what to retain, update, or forget.

In contrast, current Foundation Models for Robotics (FMRs) are largely built upon the Transformer architecture [2], which has become the dominant backbone for multi-modal foundation models due to its ability to model long-range dependencies through the self-attention mechanism. By processing the entire input sequence simultaneously, Transformers can capture complex relationships between modalities—such as vision, language, and proprioception—making them highly effective for tasks requiring semantic understanding, cross-modal alignment, and flexible reasoning. This parallelism also facilitates scalable pretraining on large datasets, a key enabler of recent advances in vision-language-action (VLA) models.

However, this powerful capacity comes with a trade-off: Transformers exhibit a quadratic computational cost with respect to input sequence length. This limitation makes them less suitable for tasks requiring long-horizon reasoning, low-latency control, or memory efficiency, all of which are crucial in real-world robotic systems. Additionally, Transformers lack an inherent mechanism for persistent memory or temporal state tracking, which restricts their performance in continuous sensorimotor tasks where maintaining a history of interaction is critical.

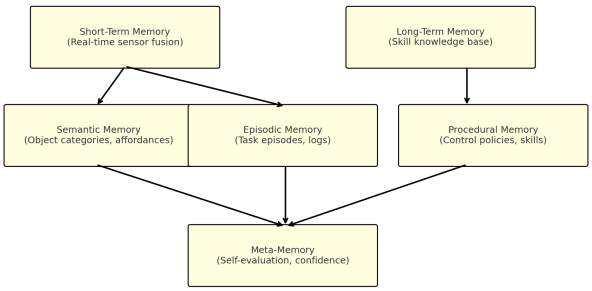

Fig. 1. Human-inspired memory architecture and its robotic analogs.

Researchers have become increasingly aware of the limitations posed by standard Transformer architectures, particularly regarding their computational inefficiency, lack of persistent memory, and inability to adapt over long horizons. In response, the *Titans* model has been proposed as a foundational framework that directly addresses these shortcomings [11]. Titans introduces a modular memory-augmented architecture that integrates components such as long-term memory modules (LMMs), surprise-based learning, and data-dependent memory updates to enable efficient, scalable, and lifelong adaptation. Notably, Titans implements an *adaptive forgetting mechanism* by prioritizing the storage of surprising information and suppressing predictable or redundant inputs. This approach avoids memory overload and supports robust knowledge consolidation over time, offering a promising path forward for building foundation models that are not only data- and compute-efficient but also capable of real-world generalization and embodied reasoning across tasks and environments.

While the Titans model marks a significant step toward integrating memory into foundation models, it remains a partial implementation of the full spectrum of human memory systems. Titans primarily addresses long-term storage and surprise-based updating, but omits crucial cognitive functions such as procedural, semantic, episodic, and meta-memory. These memory systems are not only biologically validated but also highly relevant for robotic autonomy. For instance, **procedural memory**, which in humans encodes motor skills like walking or riding a bike, would allow robots to store and reuse low-level control primitives such as grasping or tool use. **Semantic memory**, responsible for factual and conceptual knowledge, could empower robots to link object affordances and scene semantics to language commands—underpinning vision-language-action (VLA) capabilities and enabling semantic-level planning and generalization. **Episodic memory**, which allows humans to recall context-rich experiences, would let robots reflect on past events, improving task decomposition, debugging, and causal reasoning. Finally, **meta-memory**—the ability to monitor

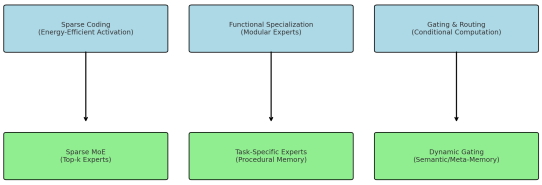

Fig. 2. Mapping brain-inspired principles to FMR mechanisms: sparse coding to Sparse MoE, functional specialization to procedural experts, and routing to memory-informed dynamic gating.

and manage one's memory—would enable self-evaluation, adaptive forgetting, and confidence estimation, critical for safe and trustworthy autonomous decision-making. By incorporating these richer memory mechanisms as illustrated in Fig.1, future FMRs could become more adaptive, explainable, and capable of lifelong learning in dynamic, open-world environments.

IV. Bringing Neural Efficiency to FMRs

The human brain is a model of computational efficiency and specialization, capable of real-time, adaptive intelligence while operating on just 20 watts of power. This remarkable capability arises from several key principles. First, the brain relies on **sparse coding**—only a small subset of neurons activate in response to any given stimulus—leading to energy-efficient, noise-robust processing [12]. Second, it exhibits **functional specialization and modularity**, with distinct brain regions (e.g., visual cortex, motor cortex, hippocampus) dedicated to different cognitive and sensorimotor tasks [13]. Importantly, **gating and routing mechanisms**—guided by attention, task demands, and internal state—determine which circuits are activated and how information flows, enabling **conditional computation**.

By contrast, most current **Foundation Models for Robotics (FMRs)** operate under monolithic architectures where all parts of the model are involved in processing every input. This leads to high computational overhead, lack of task-specific specialization, and poor scalability in energy- and time-constrained robotic systems. These models often lack mechanisms for conditional routing of information and do not adapt their computational pathways dynamically based on the input or context. As a result, even though FMRs have shown strong generalization in simulated or static tasks, their deployment in real-world, adaptive robotic systems remains constrained by latency, energy use, and brittleness.

Recognizing these limitations, researchers have proposed **Sparse Mixture of Experts (MoE)** architectures as a promising path forward. In Sparse MoE, the model consists of many specialized subnetworks (experts), but only a small number are activated per input, reducing compute while improving specialization. This approach builds upon early ideas of expert routing [14] and adaptive attention mechanisms [15], and has been successfully scaled in models like GShard [16] and Switch Transformers [17]. It is now gaining traction in robotics and multi-modal learning.

To bring neural efficiency to FMRs, future architectures must incorporate sparse expert activation, modular learning pathways, and dynamic routing mechanisms inspired by biological cognition as illustrated in Fig.2. This includes leveraging MoE to enable scalable models with input-dependent specialization, integrating attention-based or surprise-driven gating, and adopting hierarchical, brain-like modularity that supports reuse and adaptation across tasks, robots, and environments. Such pathways not only promise drastically improved efficiency, but also pave the way toward truly autonomous, self-improving robotic intelligence.

However, to achieve even greater neural-like efficiency, future architectures must go beyond current sparse MoE implementations. This includes incorporating **dynamic expert routing** influenced by uncertainty, novelty, or reward signals—mimicking top-down attention in the brain. Additionally, integrating **hierarchical MoE layers** that reflect cortical structures could enable more abstract, compositional representations. Models could also benefit from **task-aware or experience-aware gating**, where expert selection evolves based on memory and meta-cognition, similar to how humans recruit different strategies based on familiarity or context. These extensions would bring FMRs closer to the flexible, efficient, and context-sensitive intelligence found in biological systems. For instance, specific experts could be trained to represent procedural memory—encoding reusable motor skills or control primitives—allowing robots to activate the appropriate skill module based on context. In parallel, semantic memory could serve as a high-level guide for expert selection, using conceptual knowledge and affordance information to steer routing decisions. Together, these memory-informed mechanisms could foster more interpretable, adaptive, and efficient robotic intelligence.

V. Integrating Sensorimotor Feedback in FMRs

Humans and animals navigate, learn, and act in the world through the seamless integration of multiple sensory modalities. Vision allows us to perceive our environment at a distance and identify objects; hearing provides spatial and social cues; touch conveys information about texture, pressure, and temperature; and smell enables the detection of chemical signals that signal danger or opportunity. These sensory channels—each with its unique resolution, latency, and bandwidth—work in concert to guide decision-making, motor control, and learning. This sensorimotor integration is foundational for coordinated movement, object manipulation, exploration, and survival. By continuously fusing inputs from vision, hearing, proprioception, and tactile sensing, biological agents achieve a form of embodied intelligence that is both adaptive and efficient.

While foundation models for robotics (FMRs) have increasingly embraced vision-language inputs to guide behavior, robust and generalizable robotic intelligence requires the integration of a broader range of **sensorimotor feedback**, particularly **tactile and proprioceptive signals**. In contrast to vision, which offers global scene understanding, touch provides **dense, high-resolution, and spatially distributed**

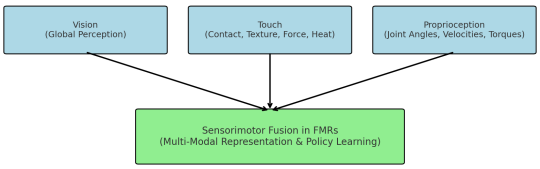

Fig. 3. Sensorimotor feedback modalities—vision, touch, and proprioception—must be fused in FMRs to support real-time, adaptive manipulation and control.

feedback along the robot's body. This tactile sensing should also incorporate **thermal information**, enabling the robot to detect temperature variations during contact—crucial for safety and material awareness. These local, contact-rich signals are essential for physical interaction, manipulation, and reactive adaptation [18].

Aware of the fact that vision alone is insufficient to capture crucial aspects of real-world manipulation, such as contact force, slippage, or texture [18], researchers have proposed vision-based tactile sensors (e.g., DIGIT [19], GelSight [20]) and tactile representation learning to enable robots to perceive detailed surface geometry and contact dynamics. When coupled with proprioception, which informs joint positions, velocities, and internal force estimates, these modalities form the basis of **reactive low-level control and adaptive physical intelligence**. However, current tactile sensors remain limited in several key ways. They are **typically localized to specific contact points**, such as fingertips or grippers, and fail to provide whole-body coverage. Most designs **do not offer temperature sensing**, a critical modality for detecting object properties, safety risks, or human contact. For truly robust physical interaction, **tactile sensing in robots should be skin-like**: distributed across the body, soft, conformable, and capable of capturing pressure, shear, and heat over large surface areas.

To fully exploit sensorimotor feedback, FMRs must move beyond vision-language-action models and adopt architectures that can process and fuse *multi-modal temporal streams*—integrating visual perception, tactile feedback, and internal body state in real time as illustrated in Fig.3. This demands not only novel data collection pipelines that capture rich contact and motion data, but also representation learning strategies that align different modalities in a shared, action-centric space. Foundation models can benefit from **cross-modal contrastive learning**, **tactile-conditioned policy learning**, and **feedback-driven action refinement** mechanisms.

Incorporating touch and proprioception into FMRs promises more reliable, compliant, and autonomous robots capable of handling uncertainty, fragility, and variability in real-world tasks. From adjusting grip based on slip, to delicately folding fabric, or securely handing over objects to humans, sensorimotor feedback is a key enabler of safe and intelligent physical interaction.

VI. Enabling FMRs with Grounded Neurocognitive Structured Reasoning

Humans acquire structured reasoning not only through abstract thinking but through years of grounded perceptual and motor experiences [21, 22, 23] . Early in development, infants begin to form causal models of the world by interacting with objects, observing outcomes, and forming sensorimotor routines. Over time, these interactions give rise to increasingly complex forms of reasoning: from spatial awareness and tool use to hierarchical goal planning and symbolic inference. Another critical element is **grounded semantic memory**, which connects abstract concepts to perceptual and sensorimotor experiences, enabling context-aware reasoning and flexible task generalization. For instance, a child learns that a 'cup' is not just an object with a label but something that affords holding liquids, is associated with drinking, and can be grasped in various ways depending on its orientation and content. A key component of this progression is **procedural memory**, which encodes reusable, composable motor routines and supports the chaining of low-level skills into complex behaviors. This progression is deeply embodied—human reasoning is anchored in bodily action, episodic memory, and multisensory feedback. Neural mechanisms supporting this development include the prefrontal cortex for executive planning, the motor cortex for action routines, and associative cortical areas that integrate sensory inputs with past experiences and conceptual knowledge.

Recent efforts in robotics and AI have seeked for such grounded neurocognitive capabilities and proposed first steps toward **Embodied Chain-of-Thought (E-CoT)** reasoning [24, 6]. These approaches extend traditional chain-of-thought reasoning—originally developed for language models—to agents that interact physically with their environment. By decomposing high-level goals into structured action steps, E-CoT enables FMRs to plan and act in a way that is more interpretable and generalizable. Recent models combine high-level symbolic planning with low-level action primitives, integrating vision-language models with trajectory prediction or affordance-based control. Some systems also incorporate feedback loops for updating plans based on real-time sensory inputs.

However, current E-CoT implementations remain far from replicating the robustness and flexibility of human grounded reasoning. They often rely on static decompositions, predefined action vocabularies, or scripted planners, and struggle to revise plans when faced with novel or uncertain environments. Unlike humans, they lack rich episodic memory for recalling past strategies, semantic grounding across modalities, or procedural fluency for fluid motor execution. Moreover, they typically do not engage in self-monitoring or introspection, making them brittle in open-ended tasks [25].

To advance toward truly grounded structured reasoning in FMRs, future research must go beyond symbolic decomposition and integrate reasoning directly with embodied experience. This includes: learning hierarchical latent action

abstractions grounded in multimodal feedback; incorporating memory architectures that blend semantic, episodic, and procedural knowledge; enabling real-time plan adaptation and failure recovery; and training models in environments that require flexible goal re-evaluation and causal inference [26, 27]. By aligning reasoning processes with embodied sensorimotor loops, FMRs can evolve toward more autonomous, explainable, and human-like intelligence.

VII. Making FMRs Self-Play in Kindergartens

Human children learn through spontaneous play, curiosity, and trial-and-error exploration [21]. In kindergarten and beyond, they experiment, manipulate objects, and invent games—not for extrinsic rewards, but for the intrinsic joy of discovery. These self-motivated activities are critical for developing fine motor skills, physical coordination, social reasoning, and causal understanding. To build Foundation Models for Robotics (FMRs) that achieve similar levels of generalization and adaptability, we must endow them with mechanisms for **self-motivated learning** and **exploratory interaction**.

In robotics, this translates into developing systems that can learn not only from expert demonstrations or task-specific supervision, but also from *self-play*, curiosity-driven exploration, and simulated environments that present a diverse range of interactive scenarios. **Simulation-based learning** plays a key role in enabling FMRs to accumulate diverse experiences required for scalable self-supervised learning [28, 29, 30]. However, most existing simulation environments are **task-specific and static** [31, 32], limiting their usefulness for open-ended skill discovery. To achieve more generalizable and adaptive behavior, FMRs should be trained in **dynamically generated simulation environments** that can be configured on the fly—e.g., through **text prompts** or structured language instructions [33]. This would allow robotic agents to encounter novel scenes, object combinations, and goals, expanding their experiential diversity.

Within these environments, FMRs can safely and efficiently explore physical dynamics, object affordances, and action outcomes across millions of episodes—much like how children acquire intuition through repetitive play. Rich simulated environments can provide randomized variations in tasks, objects, and contexts, facilitating **curriculum learning**, robustness, and transfer.

Moreover, self-motivated agents can be equipped with **intrinsic motivation signals**, such as prediction error, novelty, empowerment, or learning progress, to drive autonomous exploration. When coupled with memory and generalist policies, these mechanisms allow FMRs to acquire transferrable sensorimotor skills, adapt to new goals, and develop priors about the physical world without relying solely on human supervision. This paradigm aligns with recent advances in unsupervised and self-supervised reinforcement learning, and offers a scalable path toward autonomous skill acquisition.

Through self-play in these varied and open-ended environments, FMRs also acquire an internal **world model**—a predictive representation of the environment that enables them to simulate the outcomes of actions, reason about causality, and plan ahead. This internal model serves as the robot's embodied understanding of how the physical world behaves, supporting imagination-based learning and efficient decision-making [34]. This is essential for **lifelong learning**, where robots continually build upon prior experience, composing new skills from previously acquired ones, and refining their world model over time. Through this process, FMRs develop the ability to reuse, adapt, and generalize knowledge across diverse tasks and environments.

For open-ended multi-task learning, to enable the generalization not only horizontally to target tasks of similar complexity to the source tasks, but also to upscale to target tasks of growing complexity, the idea of decomposition and composition of actions into sequences of reusable primitives has been formalised as the motor schema theory by [35]. The theory considers that a set of action primitives might be memorised, to be retrieved and combined by the higher level to generate desired actions. The ability to compositionally combine behaviors is thought to be central to generalized intelligence in humans and a necessary component for artificial intelligent systems. A hierarchical description of actions has been proposed in neuroscience (eg [36]) and in behavioural psychology (eg. [37]).

To tackle learning multiple tasks of complexities unknown a-priori in open-ended learning [25], the robot needs to face the curse of dimensionality, as the search for policies is doubled with a search for tasks to learn, while the possible tasks space increases combinatorially. Therefore, efficient exploration is crucial. Applying *Reinforcement Learning* (RL) [38] to compositional tasks is all the more challenging as the rewards for temporally extended tasks are even more sparse. Hierarchical Reinforcement Learning (HRL) [39] has been proposed to manage long-horizon tasks. Recent works in HRL [40, 41, 42, 43, 44] have shown that learning an abstract goal representation is key to proposing semantically meaningful subgoals and to solving more complex tasks. In particular, representations that capture environment dynamics over an abstract temporal scale have been shown to provide interesting properties with regards to bounding the suboptimality of learned policies under abstract goal spaces ([41, 45, 44]), as well as efficiently handling continuous control problems. Other works [46, 47, 48] have studied various forms of spatial abstractions for goal spaces. On spatial or temporal abstractions, FMRs promise reliable representations generalisable to various real-world settings and tasks, to build upon hierarchical tasks.

However, the previously cited works in HRL have only shown their performance in a virtual environment, with a finite level of hierarchy. To tackle the increase of the task space in open-ended hierarchical learning, intrinsic motivation has been useful for robots in real-world setting and in simulation to explore both the high-dimensional task space and the policy space and to devise autonomously their learning curriculum [49, 50]. These active learning algorithms based on intrinsic motivation choose at the same time what target task to focus on, which source tasks to reuse

and how to transfer knowledge about task decomposition.

To approach the generality and flexibility of human development, FMRs should be given the capacity to learn *through play*: guided by curiosity, simulated experience, and self-generated goals, while leveraging feedback from multiple sensory modalities. Such self-driven learning loops are key to building truly adaptive, life-long learning robots capable of thriving in open-ended, dynamic environments.

VIII. Interactive Robot Learning

While FMRs are mostly black box models, the question of alignment with human representation of actions remains unaddressed, despite its importance in communication with humans about actions or tasks, and to enhance human-robot collaboration with coordination and turn taking between humans and robots. Seurin et al. [51] has explored how Natural Language can convey multiple sub-tasks by describing what the agent must accomplish, and showed that the efficiency of the communication instructions in natural language can be increased by repeated interaction between the robot and the tutor. Interaction between the robot and the tutor seems key to aligning representations. However, for FMRs that need massive data to train, they need to be complemented with self-exploration. As part of human-in-the-loop approach [52], the field of *Interactive learning* assumes that a human will be able to assist the robot in the evaluation by providing feedback, guidance and/or showing optimal actions.

Whereas reinforcement learning and supervised (or imitation) learning have been traditionally opposed we argue that both worlds are on the contrary complementary and highlight the merits of merging the two fields. While an increasing stream of machine learning works propose to combine the two paradigms [53, 54], including reinforcement learning from human feedback [55, 56, 57] including for large language models [58], most often, the agent in these works undergoes the interaction passively. We will refer to the approaches where the robot optimizes its interaction with tutors in an active way as *active imitation learning*, a term defined in [59], as an imitation learning paradigm covering cases where an observer can influence the frequency and the value of demonstrations that it is shown.

Research in developmental psychology indicates that social interaction is not only for social pleasure but can serve for learning and exploration of the environment. While most theories of infant social learning focus on how infants learn whatever and whenever the adults decide to teach them, recent findings suggest that social learning is not a passive process but that infants play an active role in collecting information and adapting their learning strategy according to their interests. Indeed, infants show a preference to learn from reliable people [60], and their attention towards adults is influenced by the adult's role in bringing new information [61]. Infants show curiosity and active contribution to social transmission of knowledge [62].

For robot learning, Nguyen [63] proposes to frame the interaction with human teachers as a reinforcement learning problem, to enable learning agents to **learn multiple parametrised tasks by devising their own learning strategy: they choose actively what do learn, when to learn and thus their own curriculum; and also what, when and whom to imitate**. Reinforcement learning is therefore not only about an agent interacting with a physical environment but also with a social environment. Using intrinsic motivation as an active learning criterion, SGIM-ACTS [64] learns several parametrised tasks by choosing its teachers and the timing of its requests, while SGIM-PB [65] uses transfer of knowledge between tasks by learning the hierarchical relationship between the parametrised tasks, showing an alignment with the human representation of task hierarchy.

Thus, active imitation learning based on intrinsic motivation can be key for FMRs to simultaneously collect data efficiently and to align its representation with humans, for efficient communication and collaboration.

IX. Conclusions

This paper explored several foundational dimensions for advancing Foundation Models for Robotics (FMRs), with an emphasis on bio-inspired principles such as memory, grounded structured reasoning, sensorimotor feedback, and self-motivated interaction-based learning. Drawing from neuroscience and developmental psychology, we outlined how robots can benefit from memory systems analogous to those of humans, how embodied chain-of-thought (E-CoT) reasoning offers a pathway toward interpretable, goal-directed robotic behavior and how interactive robot learning can be key to upscaling to tasks of growing complexity and aligning the agent's representation with humans.

While we focused on perception, reasoning, and skill acquisition, critical aspects of **reliability, safety, and trustworthiness** remain underexplored in this paper due to space limitations. These are fundamental properties for deploying FMRs in real-world environments, particularly in human-facing applications. Key open challenges include how to align robot actions with human safety norms and societal values, how to ensure robust behavior under distributional shifts or unexpected conditions, and how to make FMRs interpretable and verifiable in their decision-making.

Acknowledgment

This work was in part supported by the French Research Agency ANR, through the projects Chiron (ANR-20-IADJ-0001-01), Aristotle (ANR-21-FAI1-0009-01), and Astérix (ANR-23-EDIA-0002), and the PSPC FAIR WASTE project.

References

[1] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022.

[2] Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. 2017.

[3] Chengwei Qin et al. *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* 2023.

[4] Hans Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard Univ Press, 1988.

[5] Danny et al. Driess. “OpenVLA: Generalist foundation model for vision-language-action robotics”. In: *arXiv preprint arXiv:2312.09999* (2023).

[6] Anthony Brohan et al. “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control”. In: *arXiv preprint arXiv:2307.15818* (2023).

[7] Jiajun Wu et al. “Octo: An Open-Source Generalist Model for Robotic Policy Learning”. In: *arXiv preprint arXiv:2401.14567* (2024).

[8] D Ha et al. “Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition”. In: *arXiv preprint arXiv:2307.14535* (2023).

[9] Wei Xian et al. “Language is Not All You Need: Aligning Perception with Language Models”. In: *arXiv preprint arXiv:2302.14045* (2023).

[10] Larry R Squire. “Memory systems of the brain: A brief history and current perspective”. In: *Neurobiology of learning and memory* 82.3 (2004), pp. 171–177.

[11] Jiayuan Zhang et al. “Titans: Memory-Centric Foundation Models with Surprise-Based Learning and Adaptation”. In: *arXiv:2501.00663* (2024).

[12] Bruno A. Olshausen and David J. Field. “Emergence of simple-cell receptive field properties by learning a sparse code for natural images”. In: *Nature* 381.6583 (1996), pp. 607–609.

[13] Jerry A. Fodor. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, 1983.

[14] Noam et al. Shazeer. “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer”. In: *arXiv preprint arXiv:1701.06538* (2017).

[15] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. “Reformer: The efficient transformer”. In: *arXiv preprint arXiv:2001.04451* (2020).

[16] Dmitry et al. Lepikhin. “GShard: Scaling giant models with conditional computation and automatic sharding”. In: *arXiv preprint arXiv:2006.16668* (2020).

[17] William Fedus, Barret Zoph, and Noam Shazeer. “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity”. In: *arXiv preprint arXiv:2101.03961* (2022).

[18] Hyeongjin Lee, Yeonghoon Kim, and Hyungmin Kim. “Recent advances in tactile sensing technology for human-interactive systems: A review”. In: *Sensors* 19.19 (2019), p. 4369.

[19] Mike et al. Lambeta. “DIGIT: A Novel Design for Deep Optical Tactile Sensors”. In: *IEEE/RSJ IROS*. IEEE. 2020, pp. 1927–1934.

[20] Wenzhen Yuan, Shuang Dong, and Edward H Adelson. “GelSight: High-resolution robot tactile sensors for estimating geometry and force”. In: *Sensors* 17.12 (2017), p. 2762.

[21] Jean Piaget. *The Construction of Reality in the Child*. Basic Books, 1954.

[22] Lev S Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

[23] Andrew N Meltzoff. ““Like me” as a building block for understanding other minds”. In: *Intentions and intentionality: Foundations of social cognition* (2007), pp. 171–191.

[24] S. Liu et al. “Robotic Control via Embodied Chain-of-Thought Reasoning”. In: *arXiv preprint arXiv:2407.08693* (2024).

[25] Stephane Doncieux et al. “Open-Ended Learning: A Conceptual Framework Based on Representational Redescription”. In: *Frontiers in Neurorobotics* 12 (Sept. 2018).

[26] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, et al. “Recurrent independent mechanisms”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[27] Yilun Zhou, Rohan Shah, et al. “Inner monologue: Embodied reasoning through planning with language models”. In: *arXiv preprint arXiv:2305.14325* (2023).

[28] Josh et al. Tobin. “Domain randomization for transferring deep neural networks from simulation to the real world”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 23–30.

[29] Abhishek Gupta, Clemens Eppner, and Sergey Levine. “RoboTurk: A crowdsourcing platform for robotic skill learning through imitation”. In: *Robotics: Science and Systems (RSS)* (2021).

[30] Guillaume Duret et al. “PickSim: A dynamically configurable Gazebo pipeline for robotic manipulation”. In: *Advancing Robot Manipulation Through Open-Source Ecosystems - IEEE International Conference on Robotics and Automation (ICRA) Workshop*. 2023.

[31] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. “Jacquard: A Large Scale Dataset for Robotic Grasp Detection”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 3511–3516.

[32] Guillaume Duret et al. “FruitBin: A large-scale fruit bin picking dataset tunable over occlusion, camera pose, and scenes for 6D pose estimation”. In: *ICCV workshop* (2023).

[33] Boyuan Jiang et al. “VIMA: General robot manipulation with multimodal prompts”. In: *arXiv preprint arXiv:2210.03094* (2022).

[34] Danijar Hafner et al. “Dream to control: Learning behaviors by latent imagination”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[35] M. Arbib. “Handbook of Physiology: The Nervous System, II. Motor Control,” in: Cambridge, MA, USA: MIT Press, 1981. Chap. Perceptual structures and distributed motor control, pp. 1448–1480.

[36] Scott T. Grafton and Antonia F. de C. Hamilton. “Evidence for a distributed hierarchy of action representation in the brain”. In: *Human Movement Science* 26.4 (2007), pp. 590–616.

[37] Maria K Eckstein and Anne G E Collins. “How the Mind Creates Structure: Hierarchical Learning

of Action Sequences". In: *CogSci Conference of the Cognitive Science Society*. Vol. 43. 2021, pp. 618–624.

[38] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: an Introduction*. MIT Press, 1998.

[39] Andrew G. Barto and Sridhar Mahadevan. "Recent Advances in Hierarchical Reinforcement Learning". In: *Discrete Event Dynamic Systems* 13.1 (2003).

[40] Alexander Sasha Vezhnevets et al. "FeUdal Networks for Hierarchical Reinforcement Learning". In: *CoRR* abs/1703.01161 (2017).

[41] Ofir Nachum et al. "Near-Optimal Representation Learning for Hierarchical Reinforcement Learning". In: *ICLR*. OpenReview.net, 2019.

[42] Tianren Zhang et al. "Generating Adjacency-Constrained Subgoals in Hierarchical Reinforcement Learning". In: (June 2020).

[43] Siyuan Li et al. "Learning Subgoal Representations with Slow Dynamics". In: *International Conference on Learning Representations*. 2021.

[44] Mehdi Zadem, Sergio Mover, and Sao Mai Nguyen. "Reconciling Spatial and Temporal Abstractions for Goal Representation". In: *The Twelfth International Conference on Learning Representations*. 2024.

[45] David Abel et al. "Value Preserving State-Action Abstractions". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S Chiappa and R Calandra. Vol. 108. PMLR, 26–28 Aug 2020, pp. 1639–1650.

[46] Tejas D Kulkarni et al. "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation". In: *Advances in Neural Information Processing Systems*. Ed. by D. D. Lee et al. Vol. 29. Curran Associates, Inc., 2016, pp. 3675–3683.

[47] León Illanes et al. "Symbolic Plans as High-Level Instructions for Reinforcement Learning". In: *Proceedings of the International Conference on Automated Planning and Scheduling* 30 (2020), pp. 540–550.

[48] M. Garnelo, K. Arulkumaran, and M. Shanahan. *Towards Deep Symbolic Reinforcement Learning*. 2016.

[49] Nicolas Duminy, Sao Mai Nguyen, and Dominique Duhaut. "Learning a set of interrelated tasks by using sequences of motor policies for a strategic intrinsically motivated learner". In: *Proceedings of IEEE International Conference on Robotic Computing*. 2018.

[50] Alexandre Manoury, Sao Mai Nguyen, and Cédric Buche. "Hierarchical Affordance Discovery Using Intrinsic Motivation". In: *Proceedings of the 7th International Conference on Human-Agent Interaction*. HAI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 186–193.

[51] Mathieu Seurin. "Learning to Interact, Interacting to Learn Action-centric Reinforcement Learning". Theses. Université de Lille, Sept. 2021.

[52] Carl Orge Retzlaff et al. "Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities". In: *Journal of Artificial Intelligence Research* 79 (Jan. 2024), pp. 359–415.

[53] Bob Price and Craig Boutilier. "Implicit imitation in multiagent reinforcement learning". In: *ICML*. 1999, pp. 325–334.

[54] Andrew Y. Ng and Stuart Russell. "Algorithms for Inverse Reinforcement Learning". In: *in Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 2000, pp. 663–670.

[55] Andrea L. Thomaz and Cynthia Breazeal. "Teachable robots: Understanding human teaching behavior to build more effective robot learners". In: *Artificial Intelligence Journal* 172 (2008), pp. 716–737.

[56] W. Bradley Knox, Peter Stone, and Cynthia Breazeal. "Training a Robot via Human Feedback: A Case Study". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2013, pp. 460–470.

[57] Paul F Christiano et al. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[58] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 27730–27744.

[59] A.P. Shon, D. Verma, and Rajesh PN Rao. "Active imitation learning". In: *American Association for Artificial Intelligence*. Vol. 22. 1. 2007, p. 756.

[60] Diane Poulin-Dubois, Ivy Brooker, and Alexandra Polonia. "Infants prefer to imitate a reliable person". In: *Infant Behavior and Development* 2 (2011).

[61] Katarina Begus, Teodora Gliga, and Victoria Southgate. "Infants' preferences for native speakers are associated with an expectation of information". In: *Proceedings of the National Academy of Sciences* 113.44 (2016), pp. 12397–12402.

[62] Katarina Begus and Victoria Southgate. "Active Learning from Infancy to Childhood". In: *Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms*. Ed. by Megan M. Saylor and Patricia A. Ganea. Cham: Springer International Publishing, 2018. Chap. Curious Learners: How Infants' Motivation to Learn Shapes and Is Shaped by Infants' Interactions with the Social World, pp. 13–37.

[63] Sao Mai Nguyen. "The intrinsic motivation of reinforcement and imitation learning for sequential tasks". Accred. to supervise res. Sorbonne Université, 2024.

[64] Sao Mai Nguyen and Pierre-Yves Oudeyer. "Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner". In: *Paladyn Journal of Behavioural Robotics* 3.3 (2012).

[65] Nicolas Duminy et al. "Intrinsically Motivated Open-Ended Multi-Task Learning Using Transfer Learning to Discover Task Hierarchy". In: *Applied Sciences* 11.3 (2021).