

Toward Embodied AGI: A Review of Embodied AI and the Road Ahead

Yequan Wang Aixin Sun
tshwangyequan@gmail.com
axsun@ntu.edu.sg

Abstract

Artificial General Intelligence (AGI) is often envisioned as inherently embodied. With recent advances in robotics and foundational AI models, we stand at the threshold of a new era—one marked by increasingly generalized embodied AI systems. This paper contributes to the discourse by introducing a systematic taxonomy of Embodied AGI spanning five levels (L1–L5). We review existing research and challenges at the foundational stages (L1–L2) and outline the key components required to achieve higher-level capabilities (L3–L5). Building on these insights and existing technologies, we propose a conceptual framework for an L3+ robotic brain, offering both a technical outlook and a foundation for future exploration.

1 Introduction

Artificial General Intelligence (AGI) has attracted considerable attention in recent years (Bubeck et al. 2023). Meanwhile, Embodied AI has also seen rapid advancement (Liu et al. 2024). It is widely recognized that Embodied AI is either an essential pathway to achieving AGI—reflecting the indispensable role of the human body in cognition (Dahl 2024; Clausner and Croft 1999)—or should even form part of AGI’s definition itself (Feng et al. 2024; Tan and Jaiswal 2023). Rather than further examining the relationship between Embodied AI and AGI, we focus on *Embodied AGI*, starting from the current literature on *Embodied AI*, and explore how far it is from being truly *humanoid* and *general*. We propose a pragmatic definition of Embodied AGI as follows:

Definition 1 (Embodied AGI) *Embodied AGI is a form of Embodied AI that demonstrates human-like interaction capabilities and can successfully perform diverse, open-ended real-world tasks at a human-level proficiency.*

In this definition, Embodied AGI is framed as the intersection of AGI and Embodied AI, with an emphasis on human-like settings. To benchmark progress toward this goal, it is necessary to establish a set of criteria that clarifies the ultimate objective, assesses current capabilities, defines intermediate stages, and identifies key challenges and potential accelerators. Inspired by the levels of autonomous driving (dri 2025), we introduce a five-level roadmap for Embodied AGI (Section 2 and Figure 1), ranging from Level 1 (L1)—assisting with a limited set of elementary tasks, to Level 5

(L5)—independently performing open-ended tasks with humanoid behaviors.

We assess the capabilities of embodied AI in four core dimensions: (1) *omnimodal capabilities*: the ability to process a full spectrum of information modalities; (2) *humanoid cognitive abilities*: for nuance social comprehension and human-like learning mechanisms, including self-awareness, social connection understanding, procedural memory, and memory reconsolidation, as detailed in Section 4; (3) *real-time responsiveness*: the capability to conduct swift, accurate actions and duplex interaction; (4) *generalization*: the capability to adapt to open-ended environment and real-world tasks. The four dimensions are illustrated in Figure 2.

Based on the proposed five-level taxonomy and the four core dimensions of capability, we contextualize both recent developments and future directions. Recent developments in foundation models and embodied learning algorithms are briefly reviewed in Section 3, evaluating their current maturity. Our analysis reveals that significant gaps remain across all four dimensions in reaching L3+ Embodied AGI, placing the current state of Embodied AI development between Levels 1 and 2 (L1–L2). In Section 4, we identify the requirements across these four dimensions necessary for reaching Level 3 and beyond.

We observe that existing model architectures and widely adopted frameworks—such as Large Language Models (LLMs) (Brown et al. 2020; Meta 2024), Vision-Language Models (VLMs) (Liu et al. 2023; Agrawal et al. 2024), Vision-Language-Action (VLA) models (Kim et al. 2024; Brohan et al. 2023), and recent omnimodal approaches (Gemini 2024; Xu et al. 2025)—fall short of meeting the requirements for L3+ multimodal processing and precise real-time action execution. Furthermore, prevailing learning paradigms, including supervised and reinforcement learning (Schulman et al. 2017), remain inadequate for acquiring human-like behaviors and achieving robust generalization.

To help address these challenges, we propose a conceptual framework for L3+ Embodied AI learning in Section 5. It comprises two key components: (i) a model architecture for an advanced robotic agent, and (ii) an integrated set of learning algorithms designed to satisfy the core requirements: *omnimodal* processing, *humanoid* cognitive abilities, *real-time* responsiveness, and robust *generalization*. The proposed architecture and algorithms are illustrative examples drawn

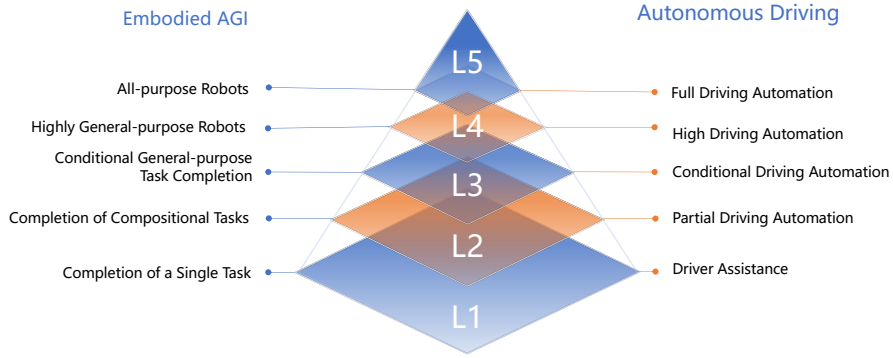


Figure 1: Roadmap of the five levels of Embodied AGI, inspired by the established levels of autonomous driving.

Table 1: Definitions of L1 ~ L5 explained. We summarize the requirements of different capabilities.

Level	Modalities	Humanoid	Real-time	Generalization	Body & Control	Autonomous Driving Analogy
L1	Partial	No	No	Unseen Environments	Robust	Simple Tasks (e.g., Speed Control)
L2	Partial	No	No	Similar Tasks	+ Responsive	Combined Tasks (e.g., Parking)
L3	Full	No	Partial	Limited Task Types	+ Sensory-complete	Complex Tasks under Human Monitoring
L4	Full	Partial	Yes	Open Tasks	+ Accurate	Broad Tasks, Humanoid Accuracy
L5	Full	Yes	Yes	Open Tasks	+ Safe	All Tasks, no Human Intervention

from current research and may be replaced by future innovations, provided they fulfill the same foundational objectives.

2 L1 ~ L5: Roadmap to Embodied AGI

Inspired by the widely accepted five levels of autonomous driving (dri 2025) and recent discussions on levels of AGI (Feng et al. 2024; Morris et al. 2023), we propose a five-stage roadmap toward Embodied AGI (L1–L5). This roadmap, summarized in Figure 1 and detailed in Table 1, defines each level based on four core dimensions (Figure 2): *modalities*, *humanoid* cognitive abilities, *real-time* responsiveness, and *generalization* capability. We also briefly outline hardware and locomotion/manipulation requirements (Gu et al. 2025) in Table 1, along with analogies to autonomous driving.

L1: Single-task completion. At this level, embodied intelligent agents (e.g., robots) reliably perform a single, well-defined task—such as object grasping—that is useful in industrial or everyday settings. While they may exhibit limited generalization to novel conditions (e.g., changes in lighting or layout), their functionality remains confined to a specific task domain. Each single task typically requires a purpose-built robot with minimal versatility, while complex goals must be manually decomposed by humans into simpler sub-tasks. This stage is analogous to early large language models focused solely on machine translation (e.g., Chinese-to-English translation) (Vaswani et al. 2017), or an L1 autonomous driving agent that handles isolated tasks like speed control and lane keeping. The robot’s physical body, at this level, must be sufficiently robust to support the execution of its target task.

L2: Compositional task completion. At Level 2, robots can handle compositional tasks by decomposing high-level human instructions into sequences of simpler actions (e.g.,

grasping followed by cutting). Their broader skill set makes them more versatile than L1 robots and reduces the need for human intervention. However, their capabilities remain bounded to predefined tasks and skill libraries, with limited generalization beyond those domains. In the LLM literature, this corresponds to a multilingual machine translation system—able to translate between many language pairs, but still confined to the translation domain. In the autonomous driving literature, this is similar to handling combined tasks with explicit decomposition logic (e.g., parking), while more complex intellectual tasks (e.g., traffic jam navigation) remain out of reach. In addition to physical robustness, the robotic body at this level must be responsive enough to support longer and more complex action sequences.

L3: Conditional general-purpose task completion. At Level 3, robots are capable of handling a wide range of task categories (e.g., grasping versus dancing), demonstrating conditional generalization across tasks, environments, and human instructions. They exhibit substantial real-time responsiveness, dynamically adapting to environmental changes or updated directives. However, while versatile and capable of multi-tasking, their performance on entirely novel or open-ended tasks is not yet reliable. Thus, L3 represents an early stage of general-purpose embodied intelligence. Supporting this level requires a robotic body with comprehensive sensory input (e.g., vision, audition, optionally touch and proprioception) and corresponding output modalities. In the context of LLMs, this stage loosely resembles pre-trained foundation models (e.g., BERT, GPT-3, LLaMA 2) equipped with multitask fine-tuning or capable of few-shot prompting (Devlin et al. 2019; Brown et al. 2020; Touvron et al. 2023). In autonomous driving, this corresponds to solving complex tasks

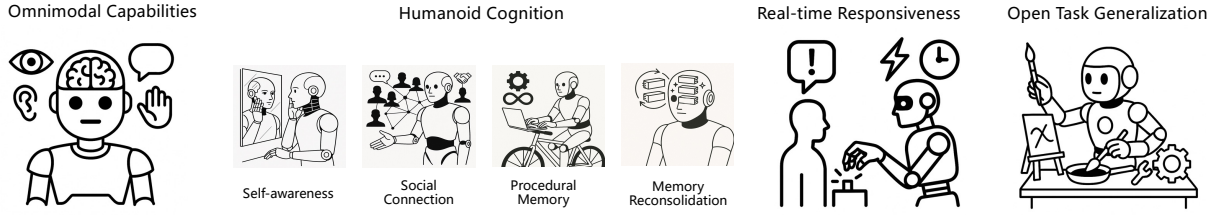


Figure 2: Illustration of four basic constituents of Embodied AGI. For a detailed discussion of the four subdomains of humanoid cognition, please see Section 4.

such as long-term highway driving and traffic navigation, albeit under human monitoring.

L4: Highly general-purpose robots. Starting from L4, robots exhibit robust generalization to a broad range of unseen tasks, marking true general-purpose capability. Such robots effectively internalize scientific laws and physical world models (Garrido et al. 2024), enabling accurate predictions and decision-making. In addition to real-time processing, they possess strong multimodal comprehension and reasoning abilities (e.g., across language, audio, and vision), ensuring sophisticated communication and interaction with humans. The robotic body is expected to be more flexible and accurate to match these advanced capabilities. For LLM analogy, L4 robots can be roughly considered as the general-purpose LLMs equipped with strong reasoning capabilities, such as o1 (Jaech et al. 2024) and DeepSeek-R1 (DeepSeek-AI et al. 2025). For autonomous driving analogy, L4 resembles solving most driving tasks at near-human accuracy, but a minimum level of human intervention is still involved.

L5: All-purpose robots. L5 represents the ultimate goal of Embodied AGI: developing genuinely all-purpose robotic agents capable of meeting a wide spectrum of human daily needs. These robots integrate a deep understanding of physical laws and human emotional and social dynamics, processing all modalities seamlessly in real-time. They exhibit distinctly human-like cognitive behaviors, including self-awareness, social connection understanding, procedural memory, and memory reconsolidation (Section 4). At this level, the robotic body should incorporate safety mechanism to prevent potential dangerous intentions from being executed. For LLM analogy, L5 corresponds to a still-emerging stage of textual AGI. In autonomous driving, it reflects a complete understanding of nuanced human needs in driving scenarios, thus fully eliminating the need for human intervention.

3 L1 ~ L2: Status Quo and Challenges

We begin with a brief literature review to assess the current status of embodied AI. Two mainstream approaches dominate the field: *end-to-end* and *plan-and-act*. End-to-end methods typically leverage Visual-Language-Action (VLA) models, directly processing visual and textual inputs to generate actions via next-token prediction (Brohan et al. 2023; Kim et al. 2024) or diffusion-based methods (Zhao et al. 2024). Conversely, plan-and-act approaches first utilize Visual-Language Models (VLMs) or Large Language Models (LLMs) to inter-

pret multimodal inputs and then perform high-level planning and task decomposition, generating intermediate control signals such as executable code (Huang et al. 2023), function calls (Kannan, Venkatesh, and Min 2024), or verbal instructions (Hu et al. 2023). Some hybrid methods integrate both paradigms through latent-space planning (AgiBot-World-Contributors et al. 2025). The notable success of LLMs (OpenAI 2023; Meta 2024) has significantly influenced foundational model development within embodied AI, promoting large-scale pre-training strategies using real-world and synthetic datasets to enhance generalization (Cheang et al. 2024; Physical-Intelligence 2025; Hu et al. 2024).

What level have we reached? Our review suggests that the capabilities required for L1 Embodied AGI are already fully or partially met by existing models. Many can reliably complete single tasks with demonstrated robustness to previously unseen environments and conditions. For example, GraspVLA (Deng et al. 2025) successfully generalizes grasping across various lighting conditions, backgrounds, distractions, and object heights. Yet, it remains specialized in grasping tasks and does not generalize beyond this domain. State-of-the-art robotic systems, such as Helix¹, not only display robust generalization within specific task types (e.g., picking diverse objects) but can also handle a wide range of dexterous indoor tasks. Such robots approach L2-level proficiency by decomposing complex human instructions into executable sub-tasks and solving them either independently or through coordinated two-bot systems.

Advancing to Level-3 (L3) requires handling substantially different task categories and exhibiting robust real-time responsiveness. Recent works, like $\pi_{0.5}$ (Physical-Intelligence 2025), partly address diverse task categories through combined pre-training (e.g., mobile and non-mobile tasks), yet their applications still largely focus on environmental generalization rather than genuine task diversity. Thus, we conclude that current Embodied AI capabilities are positioned between Levels 1 and 2 (L1–L2).

We identify four critical challenges hindering the progression of embodied AI to L3 and beyond, covering each of the four dimensions:

Lack of comprehensive joint-modal capabilities. Predominant models (e.g., VLA) typically incorporate only vision and textual language inputs, generating outputs solely in the action space. True embodied intelligence necessitates full-

¹<https://www.figure.ai/news/helix>

spectrum multimodal perception (e.g., understanding human speech with emotion and sentiment; listening to environmental audio inputs from microphone devices in addition to text console and imagery camera) and multimodal responses, including real-time acoustic speech feedback. The absence of such modalities not only severely restricts the versatility of embodied agents in application, but also prevents them from a thorough understanding of the physical world.

Insufficient humanoid cognition. Existing robots primarily focus on achieving task-specific manipulations without adequately addressing higher-level intellectual interactions or nuanced communications. Fully capable embodied agents must excel in reasoning and conversational intelligence, akin to sophisticated chatbots (Jaech et al. 2024; DeepSeek-AI et al. 2025), and demonstrate alignment with human preferences and ethical values (Ouyang et al. 2022). Ultimately, for L5, agents should exhibit distinctly humanoid cognitive behaviors and sophisticated social comprehension, which remain far beyond the reach for current learning paradigms including unsupervised, supervised, and reinforcement learning.

Limited real-time responsiveness. Most current embodied AI systems operate in a semi-duplex manner: receiving and processing instructions fully before acting, making them struggle for dynamic environments where conditions or instructions change rapidly. This limitation significantly impedes real-world deployment.

Restricted generalization. As mentioned above, recent embodied AI models have made substantial progress in generalizing across diverse environments. However, it is worth noticing that there are still a wide range of cross-environment generation scenarios that current models struggle to handle, a typical example being the invariance to spatial transformations (e.g., camera angles) (Wang et al. 2021). These issues must be addressed to reach higher levels. More importantly, inter-task generalization is still underdeveloped but essential for achieving true general-purpose capabilities (L3+).

4 L3 ~ L5: Key Constituents

In this section, we delve into the essential constituents of L3+ Embodied AGI derived from their definitions. We analyze recent advancements achieved by the research community, examine challenges that current methods face in reaching higher levels, and propose potential technical paths and design choices to bridge these gaps.

Omnimodal capabilities. A fundamental requirement of L3–L5 Embodied AGI is their “general-purpose” nature, achievable only through comprehensive omnimodal capabilities extending beyond vision and language. This is because real-world applications frequently demand an understanding of auditory cues, human speech nuances, tactile feedback, thermal perception, and more. Moreover, for L4 and beyond, mastery of these additional modalities becomes critical for acquiring and internalizing knowledge of physical laws, which is potentially the foundation of true generalization capability.

While bimodal foundational models such as visual-language (Cheang et al. 2024) and audio-language models

(Zeng et al. 2024) have been extensively explored, and tri-modal models (e.g., vision-language-audio) have garnered considerable interest recently (Gemini 2024; Xu et al. 2025), incorporating additional modalities like actions and environmental sensing for embodied agents remains largely uncharted. Moreover, current models face two critical challenges: (1) modality conflicts, which impose high demands on model capacity (Aghajanyan et al. 2023); and (2) cascading errors and alignment issues arising from modality-specific modules and heterogeneous data distributions (Tong et al. 2024). To address these issues, future models necessitate (1) *parallel* understanding-inference-generation architectures (L3+) to effectively control the time complexity imposed by model capacity, and (2) more advanced multimodal pre-training paradigms (particularly for L4+) that improves the collaboration of modality-specific modules or inherently supports multimodal comprehension.

Humanoid cognitive behaviors. Human-like cognitive behaviors are essential across all levels (L1–L5) because (1) mimicking essential learning mechanisms of human neural brain (Liu et al. 2025) potentially enhances the capability of embodied agents, and (2) a humanoid understanding of self and social connections improves the quality of human-robot interaction. Ultimately, L4+ robots should seamlessly integrate into human daily life by recognizing individual users, understanding emotional contexts, and even developing a sense of identity and social bonds (Sumers et al. 2023). We consider four capabilities being the core of achieving humanoid cognition (Figure 2):

- *Self-awareness.* As supported by cognitive science (Gallagher 2000) and philosophy (Metzinger 2004), self-awareness is the foundation of higher cognitive functions. A self-aware agent can understand its identity, temporal continuity, and objectives with greater nuance (Liang et al. 2024). This awareness should be lifelong, dynamic, and stateful—rather than statically encoded in a system prompt, as in most current LLMs.
- *Social connection understanding.* Understanding the relationships between oneself and other humans or robots—as well as relationships among others—is a higher-order cognitive capability. Such awareness helps an AI system comprehend its roles, responsibilities, and character (Chen et al. 2024), enhancing its ability to participate in role-based interactions, especially in L4+ settings. Like self-awareness, a true social connection understanding should also be lifelong, dynamic, and stateful (Fan et al. 2025).
- *Procedural memory.* Humans maintain an extendable memory of incrementally learned skills, known as procedural memory (Cavaco et al. 2004). In AI, this is related to overcoming domain shifts (Lopez-Paz and Ranzato 2017) and addressing catastrophic forgetting (Kirkpatrick et al. 2017). Agents equipped with procedural memory can accumulate and refine skills over time.
- *Memory reconsolidation.* Most current machine learning systems produce static model checkpoints after training, disallowing further learning during deployment. In contrast, humans continuously evaluate the salience of new information and update their knowledge based on time,

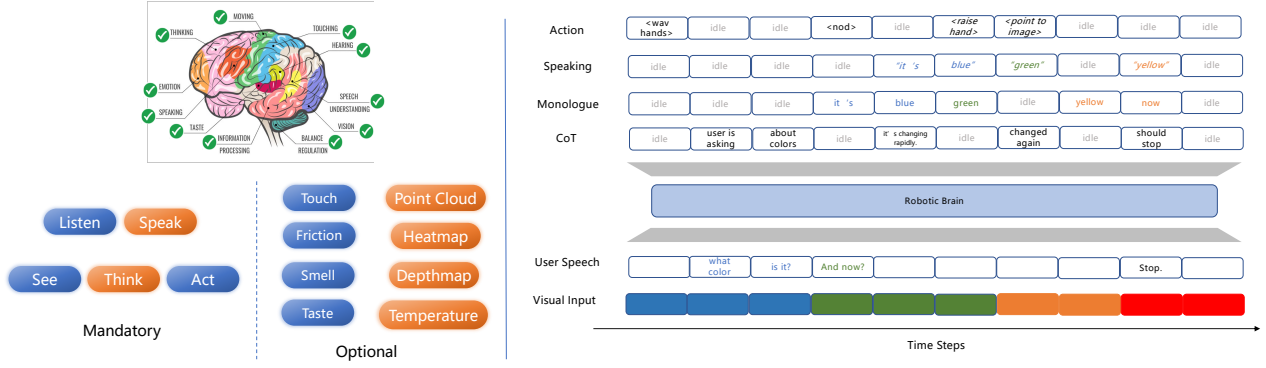


Figure 3: Omnimodal Capabilities and Illustrative Model Structure.

context, and experience—an ability known as memory reconsolidation (Alberini and LeDoux 2013). For embodied AGI, this capability is critical not only to reduce re-training overhead, but also to enable long-term adaptation and intelligence evolution.

Most of aforementioned cognition behaviors are closely connected to lifelong learning. Although recent studies have emphasized long-context learning (Wang et al. 2024a), efforts have primarily focused on extending the context window (Dao 2023) and optimizing positional encodings (Sun et al. 2023). In contrast, lifelong learning entails an unbounded temporal scope, wherein a model continuously updates its internal states and memory representations within its parameters, rather than relying on external caches (Fan et al. 2025). Human-like understanding of identity, social dynamics, and emotional contexts emerges through lifelong experiential learning supported by long-term memory. Humanoid models should therefore adopt similar lifelong learning paradigms, maintaining continuously updated internal representations of self, knowledge, and the external environment, through active, ongoing interactions.

Real-time interaction. Real-time responsiveness is essential across nearly all Embodied AI applications, especially for general-purpose agents at L3 and above, which must adapt to dynamic real-world environments and swiftly respond to rapidly-changing human instructions. Currently, real-time manipulation often imposes model-size constraints; for example, models such as GO-1 (AgiBot-World-Contributors et al. 2025) and $\pi_{0.5}$ (Physical-Intelligence 2025) employ VLA architectures restricted to under 5 billion parameters. Besides, real-time auditory and visual interactions are commonly implemented using Time Division Multiplexing (TDM) methods (Zhang et al. 2024; Wang et al. 2024b). However, these approaches encounter scalability issues when incorporating additional modalities, as computational complexity increases quadratically with sequence length (Vaswani et al. 2017). Engineering-oriented optimizations, such as those implemented in MiniCPM-o², partially alleviate this bottleneck. Nevertheless, achieving L3+ real-time performance will require new paradigms specifically designed to support gen-

uinely multiplexed, omnimodal processing.

Generalization to open-ended tasks. As discussed in Section 3, current embodied AI models demonstrate notable generalization across varied environments but struggle to generalize effectively across diverse task categories. A central limitation that hinders widely-considered unsupervised or multitask pretraining approaches from solving the problem of task generation is their insufficient internalization of *physical world laws*, which restricts their ability to accurately predict the outcomes of virtual or imagined actions. As a result, models often overfit to task-specific cues rather than uncovering underlying generalizable principles. Developing training objectives beyond simple imitation or generation—such as predictive modeling of physical interactions or causal reasoning—could significantly enhance inter-task generalization and better prepare embodied agents for open-ended, heterogeneous tasks.

5 A Conceptual Framework for L3+ Robots

In this section, we propose a conceptual framework specifically designed to meet the requirements for developing L3+ Embodied AGI, as outlined in Section 2. This framework is composed of an omnimodal model structure and a corresponding training paradigm that potentially supports the emergence of L3+ capabilities.

5.1 Model Structure

As discussed in Section 4, essential characteristics of an L3–L5 embodied AI model structure include comprehensive modality integration and native real-time interactions. Ideally, at each timestep $t + 1$, the model should generate responses conditioned on all prior information observed at timesteps $0 \dots t$. Specifically, the model jointly processes multimodal input streams, such as simultaneous audio and video, and generates multimodal outputs, including action sequences, continuous speech, internal monologues, and chain-of-thought reasoning, etc.:

²<https://github.com/OpenBMB/MiniCPM-o>

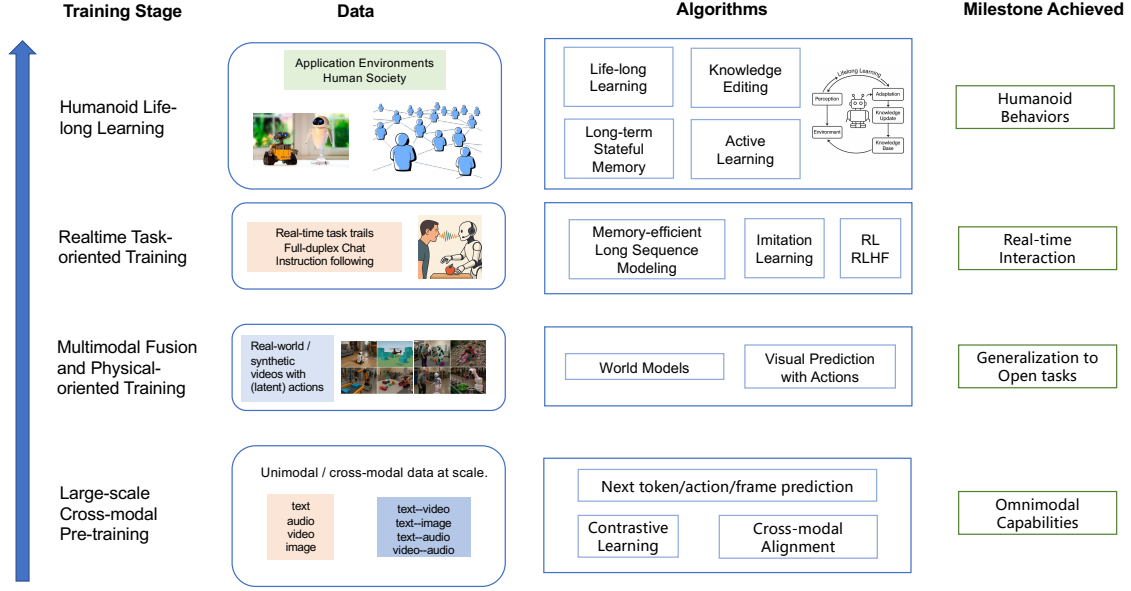


Figure 4: Conceptual Framework: Illustrative Training Paradigms.

$$y_{a_1}^{t+1}, y_{a_2}^{t+1}, \dots, y_{a_n}^{t+1} = f_{\theta}(x_{b_1}^{0 \sim t}, x_{b_2}^{0 \sim t}, \dots, x_{b_m}^{0 \sim t}). \quad (1)$$

$$a_i \in \{\text{thoughts, speech, action, mobile...}\}; \quad (2)$$

$$b_j \in \{\text{text, audio, image, video, heatmap...}\}; \quad (3)$$

An illustrative architecture is presented in Figure 3. The proposed structure supports omnimodal streaming input and output, facilitating rapid response to dynamic real-world conditions, such as changing human instructions, interruptions, environmental perturbations, and immediate feedback from previous actions. A bi-modal prototype case of such architecture is RQ-Transformer (Défossez et al. 2024).

5.2 Training Paradigm

An illustrative training paradigm is depicted in Figure 4, detailing the required data, learning algorithms, and milestone targets at each stage. The algorithms referenced are drawn from the *current AI literature* and may be replaced by future innovations serving similar objectives. The motivation and components for this paradigm are detailed as follows:

Multimodal training from scratch. We advocate training inherently multimodal models from scratch to facilitate deep cross-modal alignment and omnimodal understanding. A crucial research direction involves developing effective training stages and dataset arrangements to maximize cross-modal interactions and facilitate joint-modal comprehension.

Lifelong learning. Inspired by human cognitive behaviors, we propose moving beyond the traditional “pre-train → fine-tune → deploy” paradigm toward lifelong, continuous learning frameworks (Fan et al. 2025; Zheng et al. 2025), integrating related methodologies such as active learning (Bayer and Reuter 2024) and knowledge editing (Wang et al. 2024c) for multimodal embodied agents.

Physical-oriented training. To improve generalization across open-ended task scenarios essential for higher-level Embodied AGI, we propose to explore training paradigms oriented toward physical-world understanding. These approaches should leverage unsupervised or synthetic data at scale and incorporate explicit or implicit actions within the learning objective, allowing models to internalize causal effects and physical laws. Promising directions include outcome-prediction frameworks driven by fine-grained actions (Hu et al. 2024), and the expansion of generalized World Models (Garrido et al. 2024; Bar et al. 2024) to cover a broader domain of tasks and interaction dynamics.

6 Conclusion and Future Challenges

In this paper, we present a comprehensive review of the development of Embodied AGI by establishing a five-level taxonomy as a roadmap, benchmarking current progress, identifying critical capability gaps, and proposing a conceptual framework. We argue that this roadmap remains relevant in the long term, though advancements in robotic hardware, infrastructure, and machine learning may lead to the evolution, modification, or replacement of the proposed framework as an implementation strategy.

Our discussion is grounded in the premise that Embodied AGI should demonstrate human-like intellectual behaviors. Consequently, future challenges will not only include technical barriers but also ethical and safety considerations, as well as broader societal implications—particularly concerning the dynamics of collaboration and relationships among humans, robots, and human-robot collectives.

We hope this paper contributes valuable insights and stimulates meaningful discussion toward a more informed and constructive future for embodied general intelligence.

References

2025. The Five Steps of Automated Driving. <https://www.bosch-mobility.com/en/mobility-topics/the-five-steps-of-automated-driving/>.
- Aghajanyan, A.; Yu, L.; Conneau, A.; Hsu, W.-N.; Hambarzumyan, K.; Zhang, S.; Roller, S.; Goyal, N.; Levy, O.; and Zettlemoyer, L. 2023. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, 265–279. PMLR.
- AgiBot-World-Contributors; Bu, Q.; Cai, J.; Chen, L.; Cui, X.; Ding, Y.; Feng, S.; Gao, S.; He, X.; Huang, X.; Jiang, S.; Jiang, Y.; Jing, C.; Li, H.; Li, J.; Liu, C.; Liu, Y.; Lu, Y.; Luo, J.; Luo, P.; Mu, Y.; Niu, Y.; Pan, Y.; Pang, J.; Qiao, Y.; Ren, G.; Ruan, C.; Shan, J.; Shen, Y.; Shi, C.; Shi, M.; Shi, M.; Sima, C.; Song, J.; Wang, H.; Wang, W.; Wei, D.; Xie, C.; Xu, G.; Yan, J.; Yang, C.; Yang, L.; Yang, S.; Yao, M.; Zeng, J.; Zhang, C.; Zhang, Q.; Zhao, B.; Zhao, C.; Zhao, J.; and Zhu, J. 2025. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. *CoRR*, abs/2503.06669.
- Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D. S.; Chudnovsky, J.; Costa, D.; Monicault, B. D.; Garg, S.; Gervet, T.; Ghosh, S.; Héliou, A.; Jacob, P.; Jiang, A. Q.; Khandelwal, K.; Lacroix, T.; Lample, G.; de Las Casas, D.; Lavril, T.; Scao, T. L.; Lo, A.; Marshall, W.; Martin, L.; Mensch, A.; Muddireddy, P.; Nemychnikova, V.; Pellat, M.; von Platen, P.; Raghuraman, N.; Rozière, B.; Sablayrolles, A.; Saulnier, L.; Sauvestre, R.; Shang, W.; Soletskyi, R.; Stewart, L.; Stock, P.; Studnia, J.; Subramanian, S.; Vaze, S.; Wang, T.; and Yang, S. 2024. Pixtral 12B. *CoRR*, abs/2410.07073.
- Alberini, C. M.; and LeDoux, J. E. 2013. Memory reconsolidation. *Current Biology*, 23(17): R746–R750.
- Bar, A.; Zhou, G.; Tran, D.; Darrell, T.; and LeCun, Y. 2024. Navigation world models. *arXiv preprint arXiv:2412.03572*.
- Bayer, M.; and Reuter, C. 2024. Activellm: Large language model-based active learning for textual few-shot scenarios. *arXiv preprint arXiv:2405.10808*.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K.; Han, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, L.; Lee, T. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M. S.; Salazar, G.; Sanketi, P.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H. T.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Wu, J.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *CoRR*, abs/2307.15818.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S. M.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.
- Cavaco, S.; Anderson, S. W.; Allen, J. S.; Castro-Caldas, A.; and Damasio, H. 2004. The scope of preserved procedural memory in amnesia. *Brain*, 127(8): 1853–1867.
- Cheang, C.; Chen, G.; Jing, Y.; Kong, T.; Li, H.; Li, Y.; Liu, Y.; Wu, H.; Xu, J.; Yang, Y.; Zhang, H.; and Zhu, M. 2024. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. *CoRR*, abs/2410.06158.
- Chen, H.; Chen, H.; Yan, M.; Xu, W.; Gao, X.; Shen, W.; Quan, X.; Li, C.; Zhang, J.; Huang, F.; et al. 2024. Social-bench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*.
- Clausner, T. C.; and Croft, W. 1999. Domains and image schemas. *Cognitive linguistics*, 10: 1–32.
- Dahl, E. 2024. *Incarnation, pain, theology: a phenomenology of the body*. Northwestern University Press.
- Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; and Li, S. S. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948.
- Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037.
- Deng, S.; Yan, M.; Wei, S.; Ma, H.; Yang, Y.; Chen, J.; Zhang, Z.; Yang, T.; Zhang, X.; Cui, H.; Zhang, Z.; and Wang, H. 2025. GraspVLA: a Grasping Foundation Model Pre-trained on Billion-scale Synthetic Action Data. *CoRR*, 2505.03233.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

- Fan, S.; Huang, X.; Yao, Y.; Fang, X.; Liu, K.; Han, P.; Shang, S.; Sun, A.; and Wang, Y. 2025. If an LLM Were a Character, Would It Know Its Own Story? Evaluating Lifelong Learning in LLMs. *arXiv preprint arXiv:2503.23514*.
- Feng, T.; Jin, C.; Liu, J.; Zhu, K.; Tu, H.; Cheng, Z.; Lin, G.; and You, J. 2024. How Far Are We From AGI: Are LLMs All We Need? *arXiv preprint arXiv:2405.10313*.
- Gallagher, S. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, 4(1): 14–21.
- Garrido, Q.; Assran, M.; Ballas, N.; Bardes, A.; Najman, L.; and LeCun, Y. 2024. Learning and Leveraging World Models in Visual Representation Learning. *CoRR*, abs/2403.00504.
- Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Gu, Z.; Li, J.; Shen, W.; Yu, W.; Xie, Z.; McCrory, S.; Cheng, X.; Shamsah, A.; Griffin, R.; Liu, C. K.; et al. 2025. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *arXiv preprint arXiv:2501.02116*.
- Hu, Y.; Guo, Y.; Wang, P.; Chen, X.; Wang, Y.-J.; Zhang, J.; Sreenath, K.; Lu, C.; and Chen, J. 2024. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. *arXiv preprint arXiv:2412.14803*.
- Hu, Y.; Lin, F.; Zhang, T.; Yi, L.; and Gao, Y. 2023. Look Before You Leap: Unveiling the Power of GPT-4V in Robotic Vision-Language Planning. *CoRR*, abs/2311.17842.
- Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In Tan, J.; Toussaint, M.; and Darvish, K., eds., *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, 540–562. PMLR.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; Iftimie, A.; Karpenko, A.; Passos, A. T.; Neitz, A.; Prokofiev, A.; Wei, A.; Tam, A.; Bennett, A.; Kumar, A.; Saraiva, A.; Vallone, A.; Duberstein, A.; Kondrich, A.; Mishchenko, A.; Applebaum, A.; Jiang, A.; Nair, A.; Zoph, B.; Ghorbani, B.; Rossen, B.; Sokolowsky, B.; Barak, B.; McGrew, B.; Minaiev, B.; Hao, B.; Baker, B.; Houghton, B.; McKinzie, B.; Eastman, B.; Lugaresi, C.; Bassin, C.; Hudson, C.; Li, C. M.; de Bourcy, C.; Voss, C.; Shen, C.; Zhang, C.; Koch, C.; Orsinger, C.; Hesse, C.; Fischer, C.; Chan, C.; Roberts, D.; Kappler, D.; Levy, D.; Selsam, D.; Dohan, D.; Farhi, D.; Mely, D.; Robinson, D.; Tsipras, D.; Li, D.; Oprica, D.; Freeman, E.; Zhang, E.; Wong, E.; Proehl, E.; Cheung, E.; Mitchell, E.; Wallace, E.; Ritter, E.; Mays, E.; Wang, F.; Such, F. P.; Raso, F.; Leoni, F.; Tsimpouras, F.; Song, F.; von Lohmann, F.; Sulit, F.; Salmon, G.; Parascandolo, G.; Chabot, G.; Zhao, G.; Brockman, G.; Leclerc, G.; Salman, H.; Bao, H.; Sheng, H.; Andrin, H.; Bagherinezhad, H.; Ren, H.; Lightman, H.; Chung, H. W.; Kivlichan, I.; O’Connell, I.; Osband, I.; Gilaberte, I. C.; and Akkaya, I. 2024. OpenAI o1 System Card. *CoRR*, abs/2412.16720.
- Kannan, S. S.; Venkatesh, V. L. N.; and Min, B. 2024. SMART-LLM: Smart Multi-Agent Robot Task Planning using Large Language Models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, 12140–12147. IEEE.
- Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E. P.; Sanketi, P. R.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In Agrawal, P.; Kroemer, O.; and Burgard, W., eds., *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 2679–2713. PMLR.
- Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.
- Liang, Y.; Song, Z.; Wang, H.; and Zhang, J. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv preprint arXiv:2401.15449*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Liu, J.; Shi, X.; Nguyen, T. D.; Zhang, H.; Zhang, T.; Sun, W.; Li, Y.; Vasilakos, A. V.; Iacca, G.; Khan, A. A.; et al. 2025. Neural Brain: A Neuroscience-inspired Framework for Embodied Agents. *arXiv preprint arXiv:2505.07634*.
- Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.
- Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.
- Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. <https://ai.meta.com/blog/meta-llama-3/>.
- Metzinger, T. 2004. *Being no one: The self-model theory of subjectivity*. MIT Press.
- Morris, M. R.; Sohl-Dickstein, J.; Fiedel, N.; Warkentin, T.; Dafeo, A.; Faust, A.; Farabet, C.; and Legg, S. 2023. Levels of AGI for Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

- Physical-Intelligence. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.
- Sumers, T.; Yao, S.; Narasimhan, K.; and Griffiths, T. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.
- Sun, Y.; Dong, L.; Patra, B.; Ma, S.; Huang, S.; Benhaim, A.; Chaudhary, V.; Song, X.; and Wei, F. 2023. A Length-Extrapolatable Transformer. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, 14590–14604. Association for Computational Linguistics.
- Tan, C.; and Jaiswal, S. 2023. The path to AGI goes through embodiment. In *Proceedings of the AAAI Symposium Series*, volume 1, 104–108.
- Tong, P.; Brown, E.; Wu, P.; Woo, S.; IYER, A. J. V.; Akula, S. C.; Yang, S.; Yang, J.; Middepogu, M.; Wang, Z.; et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardaş, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, C.; Wang, R.; Mandlekar, A.; Fei-Fei, L.; Savarese, S.; and Xu, D. 2021. Generalization Through Hand-Eye Coordination: An Action Space for Learning Spatially-Invariant Visuomotor Control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, 8913–8920. IEEE.
- Wang, H.; Shi, H.; Tan, S.; Qin, W.; Wang, W.; Zhang, T.; Nambi, A.; Ganu, T.; and Wang, H. 2024a. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*.
- Wang, P.; Lu, S.; Tang, Y.; Yan, S.; Xiong, Y.; and Xia, W. 2024b. A Full-duplex Speech Dialogue Scheme Based On Large Language Models. *CoRR*, abs/2405.19487.
- Wang, S.; Zhu, Y.; Liu, H.; Zheng, Z.; Chen, C.; and Li, J. 2024c. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3): 1–37.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Zeng, A.; Du, Z.; Liu, M.; Zhang, L.; Jiang, S.; Dong, Y.; and Tang, J. 2024. Scaling Speech-Text Pre-training with Synthetic Interleaved Data. *CoRR*, abs/2411.17607.
- Zhang, X.; Chen, Y.; Hu, S.; Han, X.; Xu, Z.; Xu, Y.; Zhao, W.; Sun, M.; and Liu, Z. 2024. Beyond the turn-based game: Enabling real-time conversations with duplex models. *arXiv preprint arXiv:2406.15718*.
- Zhao, T. Z.; Tompson, J.; Driess, D.; Florence, P.; Ghasemipour, S. K. S.; Finn, C.; and Wahid, A. 2024. ALOHA Unleashed: A Simple Recipe for Robot Dexterity. In Agrawal, P.; Kroemer, O.; and Burgard, W., eds., *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 1910–1924. PMLR.
- Zheng, J.; Shi, C.; Cai, X.; Li, Q.; Zhang, D.; Li, C.; Yu, D.; and Ma, Q. 2025. Lifelong Learning of Large Language Model based Agents: A Roadmap. *arXiv preprint arXiv:2501.07278*.