Agent AI Towards a Holistic Intelligence

Qiuyuan Huang[⋆ℭ▶], **Naoki Wake**[⋆ℜ▶◊], **Bidipta Sarkar**[§†], **Zane Durante**[§†],

Ran Gong[♮†], **Rohan Taori**[§†], **Yusuke Noda**[Ↄ], **Demetri Terzopoulos**[♮],

Noboru Kuno[◁], **Ade Famoti**[◁], **Ashley Llorens**[◁], **John Langford**[Γ],

Hoi Vo[Ↄ‡], **Li Fei-Fei**[§‡], **Katsu Ikeuchi**[ℜ‡], **Jianfeng Gao**[ℭ‡]

[ℭ]Microsoft Research Core, Redmond; [ℜ]Microsoft Applied Robotics Research, Redmond;
[§]Stanford University; [♮]University of California, Los Angeles;
[Ↄ]Microsoft Gaming US; [◁]MSR Accelerator; [Γ]MSR AI Frontiers, Newyork

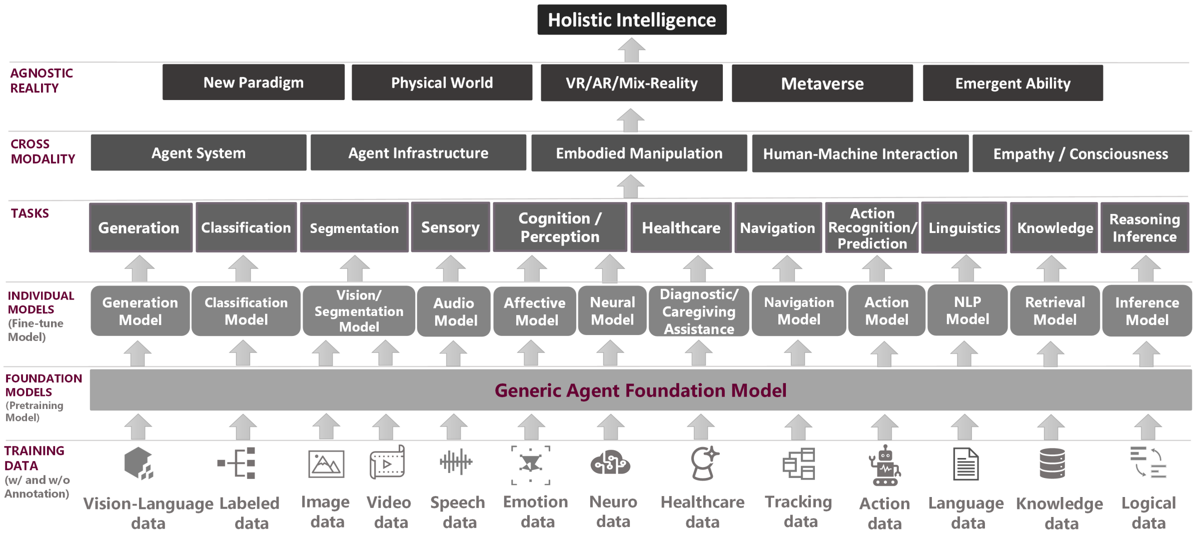


Figure 1: Overview of an Agent AI system. This system is applicable across multiple domains and provides a foundation model for interactive manipulation and embodied operations. Agent AI operates in both physical and virtual worlds by leveraging cross-modal data that is acquired through interactions between diverse environments. Agent AI offers a promising approach to unify a broad range of applications and capabilities within infrastructure and system. Furthermore, it is emerging as a promising pathway towards Holistic Intelligence (HI).

Abstract

Recent advancements in large foundation models have remarkably enhanced our understanding of sensory information in open-world environments. In leveraging the power of foundation models, it is crucial for AI research to pivot away from excessive reductionism and toward an emphasis on systems that function as cohesive wholes. Specifically, we emphasize developing Agent AI—an embodied system that integrates large foundation models into agent actions. The emerging field of Agent AI spans a wide range of existing embodied and agent-based multimodal interactions, including robotics, gaming, and healthcare systems, etc. In this paper, we propose a novel large action model to achieve embodied intelligent behavior, the Agent Foundation Model. On top of this idea, we discuss how agent AI exhibits remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. Furthermore, we discuss the potential of Agent AI from an interdisciplinary perspective, underscoring AI cognition and consciousness within scientific discourse. We believe that those discussions serve as a basis for future research directions and encourage broader societal engagement.

[⋆]Equal Contribution. [▶]Project Lead. [‡]Equal Advisor. [◊]Corresponding Author. [†]Work done while interning or researching part-time at Microsoft Research, Redmond.

1 Introduction

Artificial Intelligence (AI) was historically defined at the 1956 Dartmouth Conference as artificial life forms capable of collecting information from their environment and taking effective actions within it. Minsky's group at MIT developed a robotic system in 1970, known as the "Copy Demo," that observed "blocks world" scenes and successfully reconstructed the observed polyhedral block structures (Winston, 1972). The system, comprising observation, planning, and manipulation modules, demonstrated that each of these subproblems was highly challenging and necessitated further research. Consequently, the field of AI fragmented into specialized subfields. While these subfields have made significant progress independently, this over-reductionism has blurred the overarching goals of AI research.

To advance beyond the current state towards more sophisticated AI, we emphasize the importance of embracing the holistic philosophy of Aristotle, which underscores the integration of components to surpass the sum of its parts. Recent advancements in Large Language Models (LLMs) and Visual Language Models (VLMs) have shown great potential in recognizing language and images in an open-world context (OpenAI, 2023). For example, the advanced semantic processing of LLMs has been utilized to decompose human instructions into high-level tasks for robots (Wake et al., 2023c,d). However, these existing multimodal foundation models, even for GPT-4V(ision), still face a challenge in achieving fine-grained manipulation that necessitates action prediction. Therefore, a new embodied Agent Foundation Model was proposed (Durante et al., 2024b) which integrates language proficiency, visual cognition, context memory, intuitive reasoning, and can predict the embodied actions with adaptability. This is the first study that pretrains a foundation model for the development of general-purpose AI agents by using embodied data collected from robotics, gaming, and healthcare tasks.

An embodied agent is conceptualized as an interactive system that communicates with humans and interacts with environments through its perceptual capabilities, employing actions aligning with human intents. This is the reason why we consider the advance of large embodied foundation models as a significant contribution to Agent AI, enabling systems to parse and infer human intent from various domain information, actions, natural-language instructions and multimodal contexts. Moreover, actively leveraging action-based large foundation models makes our approach unique for developing integrated AI systems.

Building upon the Agent AI framework, we believe that the AI community will steadily accumulate insights and knowledge essential for transitioning from AI models used for passive, structured tasks to those capable of dynamic, interactive roles in complex environments. This is a critical step towards the development of Artificial General Intelligence (AGI) (Fig. 1). In this paper, we analyze a new architecture for Agent AI systems, alongside a review of recent literature in Agent AI domains including robotics, gaming, and healthcare. Furthermore, we explore the cognitive aspects of Agent AI and introduce research areas impacted by Agent AI to engage a broader community of researchers and actively promote its development. Finally, we discuss future research directions, including the ethical challenges that need to be addressed. Through these discussions, we aim to illustrate how the development of these technologies is bringing AI agents closer to AGI, holistic intelligence.

2 Agent AI Paradigm

2.1 Agent AI fundamentals

We define Agent AI as an *intelligent agent capable of autonomously executing appropriate and contextually relevant actions based on sensory input, whether in a physical, virtual, or mixed-reality environment.* Agent AI represents a new paradigm that sheds light on embodied intelligence, emphasizing the importance of an integrated approach for interactive agents in complex dynamics. This approach is motivated by the belief that intelligence arises from the intricate interplay between learning, memory, action, perception, planning, and cognition (Fig. 2).

Learning. Agent AI can adapt to new environments by acquiring new knowledge and updating its skills. To this end, the agent needs to observe its environment, understand the impact of its actions on that environment, and learn from human demonstrations (Wake et al., 2021). For instance, by employing reinforcement learning (RL) techniques or supervised learning from human demonstrations (e.g., imitation learning (IL), behavior cloning), the agent can progressively improve its behavior.

Memory. Long-term memory enables the Agent to remember specific operations adaptable to the environment or user preference. In contrast, short-term memory pertains to the history of actions taken and

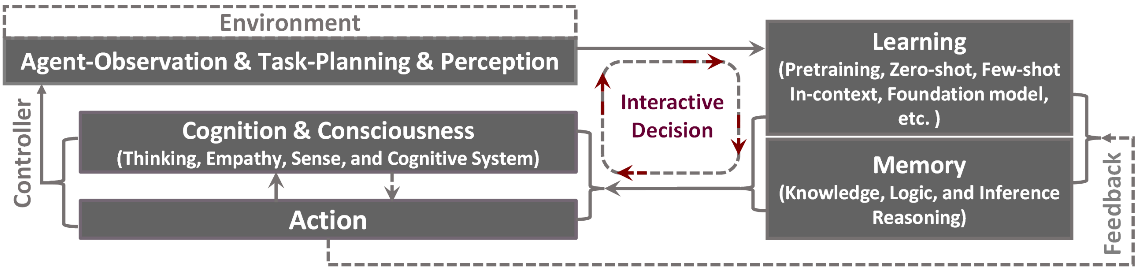

Environment
Agent-Observation & Task-Planning & Perception
Controller
Cognition & Consciousness
(Thinking, Empathy, Sense, and Cognitive System)
Action
Interactive
Decision
Learning
(Pretraining, Zero-shot, Few-shot
In-context, Foundation model,
etc.)
Memory
(Knowledge, Logic, and Inference
Reasoning)
Feedback

Figure 2: An Agent AI paradigm for supporting embodied multi-modal generalist agent systems. There are five main modules as shown: (1) Agent in Environment and Perception with task-planning and observation, (2) Agent Learning, (3) Memory, (4) Action, and (5) Cognition and Consciousness. We believe that the cohesive integration of these components facilitates the development of a holistic intelligence. A key distinction of our approach from some prior interactive strategies is that, after training, the agent's actions will directly influence task planning without the need for receiving feedback from the environment to plan its subsequent actions as the previous interacive paradigm.

perceptions observed during an operation. Short-term memory enables the system to replan and consider next-step actions based on history.

Action. The actions of Agent AI do not necessarily have to be physical actions in the real world. Depending on the definition of the environment, actions may include interactions in virtual reality (VR) environments or speech directed at humans. A suitable action is selected through a cognitive process from learned skills, based on memory. Additionally, real-world operations often cannot be completed in one shot and thus require multi-round interactions between humans or the environment and the agent. This interaction is also orchestrated by a cognitive process and memory (e.g., conversation history).

Perception. Like humans, robust and multimodal perception is crucial for agents to understand their environment. Visual perception is one of the most important abilities, enabling the agent to comprehend the world, e.g., images, videos, gameplay. Similarly, audio perception is crucial for understanding human intent.

Planning. Planning is an important aspect of long-range tasks, such as a robot manipulating objects in an environment for a specific purpose. The planning strategy typically depends on the goal of the task. Goal-oriented planning enables flexible operation that adapts to uncertainties due to any external and internal disturbances.

Cognitive Aspects. Agent AI focuses not only on the performance of individual components but also on the utility of the system as a whole. Consider a scenario where a robot, right after being unboxed, begins to communicate with a non-expert user and swiftly adapts to carry out domestic tasks within the user's home setting. Realizing such a system is challenging and requires a mechanism that orchestrates each Agent AI components. This orchestration functionality is referred to as the cognitive aspect of Agent AI.

2.2 Agent AI Consciousness

Agent AI can go beyond a simple component orchestration and potentially entail a type of "consciousness." In recent challenging attempts to find consciousness in AI based on neuroscientific insights, neuroscientists have discussed *Agency* and *Embodiment* as indicators of consciousness (Butlin et al., 2023). Agency refers to the capacity to learn from feedback, make decisions to pursue goals, and adapt to conflicting objectives. It indicates a system's characteristic of attempting to achieve goals through interaction with its environment. Embodiment involves understanding and utilizing the relationship between actions and feedback from the environment to affect perception or control. It emphasizes comprehending how one's body and the surrounding environment can be leveraged in cognitive processes.

Our Agent AI predicts optimal actions based on language (i.e., textual instructions), sensory inputs, and action history, fulfilling Agency by generating goal-directed actions. It also learns from the relationship between its actions and environmental outcomes, fulfilling the principle of Embodiment. Thus, we can potentially quantify aspects of Agent AI's consciousness, suggesting its potential across disciplines like neuroscience, biology, physics, biological physics, cognitive science, medical health, and moral philosophy.

There are various approaches to developing Agent AI. In Section 4, we will introduce a specific example of Agent AI. In Section 6, we will discuss the main challenges and necessary actions, including ethical concerns in Agent AI research.

3 Agent Foundation Model

Agent AI systems that interact with the environment, with humans, and amongst other agents. We consider agent-environment interactions as encompassing a broader scope than embodied agents. For instance, ambient intelligence systems, which despite not only having a physical embodiment, can be embedded into and interact with their environment. The advancement of agent systems that interact with humans is another area of keen interest for this area. We strongly believe that multimodal interactions between humans and agents, extending beyond high-level intention instructions, is a promising area of research and future direction for low-level fine-grained actions manipulation with human-agent interactions. We are also interested indeveloping systems for effective agent to agent communication and efficient collaboration within multi-agent infrastructures and exploring new agent paradigm and agent learning Strategy.

In this section, we provide an overview of Agent AI system that leverages foundation models with the latest machine-learning technologies. The system is composed of three components: i) Interactive agent transformer, ii) Agent foundation model learning strategy with RL and IL, and iii) self improvement.

3.1 Agent Transformer

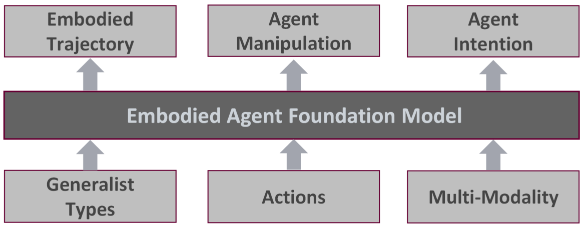


Figure 3: Overview of an interactive agent foundation model framework. The transformer is designed to process multimodal information that conveys various levels of abstraction. This approach facilitates a comprehensive understanding of the context, thus enhancing coherent actions. Through learning across a variety of task domains and applications.

We analyze a transformer-based multimodal encoder (Fig. 3) that enables an interactive agent to take actions based on multimodal information. This model is initialized with three pre-trained submodules, namely, the visual module, the agent action module and the language module.

To facilitate cross-modal information sharing, (Durante et al., 2024b) foundation model allows the agent to predict actions (or action tokens) to complete the embodied tasks in robot, gaming, and interactive healthcare domains. The model also feed diverse historical data into the transformer model including but not limited to previous low-level fine-grained actions (agent information), video/images, audio, language, or high-level instruction, as context during pre-training. As a result, for any given time step, it can predict low-level manipulation (action) tokens, general agent types (e.g., TypeChat in gaming), or high-level instructions (e.g., agent intention). Moreover, the unified transformer can also produce high-level instructions based on text prompts, visual context, and previous actions. This approach allows the model to take into account both the current context and the history of interactions, making it able to respond more accurately to the task at hand.

3.2 Agent Learning Strategy

Reinforcement Learning (RL). To learn the optimal relationship between states and actions based on rewards (or penalties) received as a result of its actions, we can use reinforcement learning. RL is a highly scalable framework that has been applied to numerous applications, including robotics. For many applications, it is challenging or costly to collect human demonstrations, such as learning policies in automatically generated virtual environments. RL is particularly effective in these scenarios, exemplified by the actor-critic algorithm PPO (Schulman et al., 2017). Additionally, RL technology can be applied to model human-AI interactions, which is a crucial aspect of interactive Agent AI. For instance, agents can be trained via RL from human feedback (RLHF) (Ouyang et al., 2022), allowing humans to choose desired responses without hand-engineering rewards.

Imitation Learning (IL). IL seeks to leverage demonstration data to mimic the actions of human experts. For example, in robotics, one of the major frameworks based on IL is Behavioral Cloning (BC). BC is an approach where a robot is trained to mimic the actions of an human expert by directly copying them. In this approach, the expert's actions in performing specific tasks are recorded, and the robot is trained to replicate these actions in similar situations. Recent BC-based methods often incorporate technologies from LLM/VLMs, enabling more advanced end-to-end models. For example, Brohan et al. proposed RT-1 (Brohan et al., 2022) and RT-2 (Brohan et al., 2023), transformer-based models that output an action sequence for a robot's base and arm, taking a series of images and language as input. These models are reported to show high generalization performance as the result of training on a large amounts of demonstration data.

Traditional RGB. Learning intelligent agent behavior leveraging image inputs has been of interest for many years (Mnih et al., 2015). The inherent challenge of using RGB input is the curse of dimensionality. To solve this problem, researchers either use more data (Jang et al., 2022; Ha et al., 2023) or introduce inductive biases into the model design to improve sample efficiency. In particular, authors incorporate 3D structures into the model architecture for manipulations (Zeng et al., 2021; Shridhar et al., 2023; Goyal et al., 2023; James and Davison, 2022). For robot navigation, authors (Chaplot et al., 2020a,b) leverage maps as a representation. Maps can either be learned from a neural network aggregating all previous RGB inputs or through 3D reconstruction methods such as Neural Radiance Fields (Rosinol et al., 2022).

3.3 Optimization in the Agent System

The optimization of agent systems can be divided into spatial and temporal aspects. Spatial optimization considers how agents operate within a physical space to execute tasks. This includes inter-robot coordination, resource allocation, and keeping an organized space. In order to effectively optimize agent AI systems, especially systems with large numbers of agents acting in parallel, previous works have focused on using large batch reinforcement learning (Shacklett et al., 2023). Since datasets of multi-agent interactions for specific tasks are rare, self-play reinforcement learning enables a team of agents to improve over time. However, this may also lead to very brittle agents that can only work under self-play and not with humans or other independent agents since they over-fit to the self-play training paradigm. To address this issue, we can instead discover a diverse set of conventions (Cui et al., 2023; Sarkar et al., 2023), and train an agent that is aware of a wide range of conventions. Foundation models can further help to establish conventions with humans or other independent agents, enabling smooth coordination with new agents (Gong et al., 2023).

Temporal optimization, on the other hand, focuses on how agents execute tasks over time. This encompasses task scheduling, sequencing, and timeline efficiency. For instance, optimizing the trajectory of a robot's arm is an example of efficiently optimizing movement between consecutive tasks (Zhou et al., 2023b). At the level of task scheduling, methods like LLM-DP (Dagan et al., 2023) and ReAct (Yao et al., 2023) have been proposed to solve efficient task planning by incorporating environmental factors interactively.

3.4 Self Improvement for Transformers

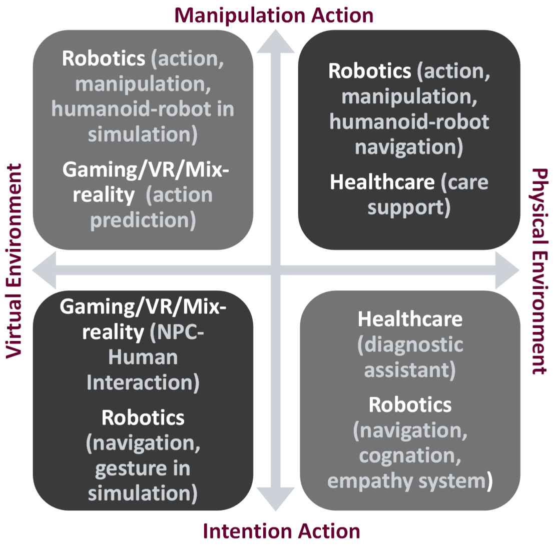
Currently, foundation model based AI agents have the capacity to learn from multiple different data sources, which allow for more flexible sources for data for training. Two key consequences of this are that (1) user and human-based interaction data can be used to further refine and improve the agent and (2) existing foundation models and model artifacts can be used to generate training data. We discuss each of these in more detail in the following sections, but we note that since current AI Agents are largely tied to existing pretrained foundation models, they generally do not learn from continuous interaction with their environments. We think this is an exciting future direction, and initial work by Bousmalis et al. has shown that self-improving agents for robotic control are able to continuous learn and improve through environmental interactions without supervision (Durante et al., 2024b; Bousmalis et al., 2023).

Furthermore, the iterative learning process can leverage human feedback (Gong et al., 2023). For example, in the context of robot teaching, Agent AI understands what it needs to do from multimodal instructions provided by humans (Wake et al., 2021). Based on these instructions, it generates images or scenes and makes them operable in a virtual world. This process is repeated by utilizing user feedback, allowing Agent AI to gradually improve and adapt itself to the environment.

4 Agent AI Categorization

Agent AI aims to develop agents that can adeptly navigate and interact with a changing world. These agents are designed to learn and solve complex tasks through direct engagement with their environment. The field has been propelled forward by significant advancements in the development of general-purpose foundation models, leading to superhuman achievements in various AI domains previously deemed challenging. These developments have significantly boosted the capabilities of embodied AI. Researchers are now rapidly advancing towards creating intelligent agents that can perceive their surroundings, engage in natural language dialogue, understand and respond to auditory inputs, navigate and manipulate their environment to achieve objectives, and reason about the long-term outcomes of their actions. We are interested in particular with submissions that focus on the multimodal aspects of embodied AI systems and develop novel methods for synthesizing meaningful agent outputs from multi-sensory inputs.

4.1 Embodied Agent Categorization



Manipulation Action
Robotics (action, manipulation, humanoid-robot in simulation)
Gaming/VR/Mix-reality (action prediction)
Robotics (action, manipulation, humanoid-robot navigation)
Healthcare (care support)
Virtual Environment
Physical Environment
Gaming/VR/Mix-reality (NPC-Human Interaction)
Robotics (navigation, gesture in simulation)
Healthcare (diagnostic assistant)
Robotics (navigation, cognation, empathy system)
Intention Action

Figure 4: Overview of the two axes for agents spaces. Embodied Agent AI is classified according to the extent to which it involves low-level fine action manipulations, which we refer to as "manipulation actions" (e.g., action prediction) in an environment, whether real or virtual. In contrast, an agent's actions may primarily aim at high-level information transmission for a robot or human's intent instruction, which we refer to as "intention action" (e.g., general task planning). An agent's environment can be broadly categorized based on whether it is the physical world or a virtual one. According to this, we divide embodied and interactive Agent into main four categories.

Agent AI refers to AI systems that integrate Large foundation models. Consequently, a number of recent AI systems that are based on LLM/VLMs can be associated with Agent AI subcategories. Specifically, we categorize Agent AI based by the types of agent actions and their environments, as illustrated in Fig. 4. Therefore, Agent AI can be broadly grouped into four categories. This section reviews related research (Durante et al., 2024a) and organizes them according to these categories. We also expand on systems combining both intention and manipulation agents in Appendix A.

4.1.1 Manipulation action in physical environments

Agents in this category are intended to work in the physical world, with robotics applications being the typical example (Ahn et al., 2022a; Huang et al., 2022b; Liang et al., 2022; Driess et al., 2023; Brohan et al., 2023). Training agents for physical manipulation in an end-to-end manner is typically challenging due to the significant costs associated with collecting a large amount of data for training. Consequently, recent trends have shifted towards solving higher-order task plans with large foundation model and integrating these with lower-level controllers that are trained using conventional methods RT-1 (Brohan et al., 2022) and RT-2 (Brohan et al., 2023) and Agent foundation model (Durante et al., 2024b).

4.1.2 Manipulation action in virtual environments

This type of agent utilizes a virtual simulated environment. In the robotics domain, the main objective is to train Agent AI through trial-and-error for tasks where physical trials are impractical or risky, including the ability to predict user actions and devise plans for tasks within specific constraints (Ahn et al., 2022b; Brohan et al., 2023; Durante et al., 2024b; Gong et al., 2023). In the case of gaming agents, the goal is not to eventually transition to the physical world, but the learning within the simulation environment itself is the main objective (Park et al., 2023c; Wang et al., 2023b,e; Baker et al., 2022).

There have also been a number of works that demonstrate the ability of general-purpose visually-aligned large language models trained on large-scale text, image, and video data to serve as a foundation for creating multi-modal agents that are embodied and can act in various environments (Baker et al., 2022; Driess et al., 2023; Brohan et al., 2023; Durante et al., 2024b). Typically, research on these agents involves simulation platforms for object recognition (Kolve et al., 2017; Wang et al., 2023d; Mees et al., 2022; Yang et al., 2023a; Ehsani et al., 2021; Szot et al., 2021; Puig et al., 2018; Carroll et al., 2019; Li et al., 2021; Srivastava et al., 2022; Mittal et al., 2023; Zhong et al., 2023; Liu and Negrut, 2021; Saito et al., 2023; Huang et al., 2022a).

4.1.3 Intentional action in physical environment

A typical example of interactive agents in this category is found in the healthcare domain, such as applications in diagnostics and knowledge retrieval (Lee et al., 2023; Peng et al., 2023). In a similar context, several works have developed empathy-aware agents for engaging dialogue and human-machine interactions (Chen et al., 2021; Mao et al., 2022; Wake et al., 2023a; Savva et al., 2019; Puig et al., 2023; Huang et al., 2018). In other cases, Agent AI's focus on knowledge and logical reasoning involves integrating implicit and explicit knowledge sources. This integration enables more accurate and contextually appropriate responses (Brown et al., 2020; OpenAI, 2023; Lewis et al., 2020; Peng et al., 2023; Gao et al., 2022; Marcus and Davis, 2019; Gao et al., 2020; Wang et al., 2023a; Chen et al., 2020; Park et al., 2023a; Li et al., 2023b).

4.1.4 Intentional action in virtual environment

Studies on Agent AI in this category have highlighted the utility for the creation of interactive content in gaming and both VR and XR (Chen et al., 2021; Mao et al., 2022; Huang et al., 2023b). Agent navigation following instruacion is also a representative task that falls in this category (Tsoi et al., 2022; Deitke et al., 2020). Similar to gaming agents for intentional action, this type of Agent AI has shown super-human performance in specific games (Meta Fundamental AI Research et al., 2022; Yao et al., 2023). Recent robotics research also leverages LLMs to perform task planning (Ahn et al., 2022a; Huang et al., 2022b; Liang et al., 2022) by decomposing natural language instruction into a sequence of subtasks, either in the natural language form or in Python code, then using a low-level controller to execute these subtasks.

4.2 Multimodel Agent Categorization (Non-Embodied)

These categories of Agents emphasize the importance of using multimodal information to take beneficial non-embodied from their respective aspects. This indicates the necessity for agents to possess high recognition capabilities for both language and vision, thereby strongly suggesting the effectiveness of leveraging large fondation models. MUltimodel Agent have shown significant utility across a variety of tasks. The advancements in large-scale foundational models and interactive artificial intelligence have opened up novel capabilities for multimodel agent. A number of works leverage multi-model agents to perform task planning (Huang et al., 2022a; Wang et al., 2023b; Yao et al., 2023; Li et al., 2023b), and leverage the large multimodels' large internet-scale domain knowledge and zero-shot planning abilities to perform agentic tasks like planning and reasoning. Additionally, (Huang et al., 2022b), (Liang et al., 2022), and (Wang et al., 2023e) also incorporate environmental feedback to improve task performance.

Nevertheless, for agent AI to be genuinely beneficial, they must offer intuitive interaction experiences and adapt to a wide array of environments, contexts, and modalities. To promote research in this area, we proposed a broad range of categorization relevant for multimodal agents without embodied action including (Gui et al., 2022a; Park et al., 2023b), but not limited to Simulation and environments agents (Puig et al., 2018)), generative agents (Huang et al., 2023b), knowledge and logical inference agents (Lewis et al., 2020; Peng et al., 2023; Wang et al., 2023a; Gui et al., 2022b), emotion agent (Chen et al., 2021), Neuro-symbolic agents (Chen et al., 2020), and agents for traditional multimodal tasks, multimodal agent systems and infrastructure, and applications of multimodal agents.

5 Agent AI Application Tasks

In Section 4, we categorized existing research within the realm of Agent AI. To offer a tangible understanding of its applications, we introduce four mission-critical domains where Agent AI can have a major impact.

5.1 Robotics

Robots are representative agents that necessitate effective interaction with their environment. In this section, we introduce key elements essential for efficient robotic operation, review research topics where the latest large foundation models have been applied, and share insights from recent studies.

Multimodal Systems. Recent research focuses on developing end-to-end systems incorporating large foundation model technologies as encoders for input information, guiding robotic actions based on linguistic instructions and visual cues (Huang et al., 2018; Jiang et al., 2022; Brohan et al., 2023, 2022; Li et al., 2023f; Ahn et al., 2022b; Shah et al., 2023b; Li et al., 2023c).

Task Planning and Skill Training. Advanced language processing abilities of LLMs interpret instructions and decompose them into robot action steps, advancing task planning technologies (Ni et al., 2023; Li et al., 2023a; Parakh et al., 2023; Wake et al., 2023b). For skill training, large foundation models are used for designing reward functions (Yu et al., 2023; Katara et al., 2023; Ma et al., 2023), generating data for policy learning (Kumar et al., 2023; Du et al., 2023), or as part of a reward function (Sontakke et al., 2023).

On-site Optimization. This involves dynamically adapting and refining robotic skills by integrating task plans with real-time environmental data (Ahn et al., 2022b; Zhou et al., 2023b; Raman et al., 2023; Chen et al., 2021). Strategies seek to achieve environment-grounded robot execution by adjusting the robot's actions at the task plan or controller level.

Conversation Agents. LLMs contribute to natural, context-sensitive interactions with humans in conversational robots (Ye et al., 2023; Wake et al., 2023d). They process and generate responses that mimic human conversation and estimate conceptual (Hensel et al., 2023; Teshima et al., 2022) and emotional attributes (Zhao et al., 2023; Yang et al., 2023b; Wake et al., 2023a) of utterances.

Navigation Agents. Robot navigation focuses on core aspects such as map-based path planning and SLAM (Guimarães et al., 2016). Recent work enables robots to navigate in challenging environments using object names (Chaplot et al., 2020a; Batra et al., 2020; Gervet et al., 2023; Ramakrishnan et al., 2022; Zhang et al., 2021) or zero-shot object navigation (Gadre et al., 2023; Dorbala et al., 2023; Cai et al., 2023). Vision-Language Navigation (VLN) interprets sentences for navigation in unseen environments (Anderson et al., 2018; Shah et al., 2023a; Zhou et al., 2023a; Dorbala et al., 2022; Liang et al., 2023; Huang et al., 2023a). VLN interprets sentences rather than object names, it requires a higher functionality to parse input text (Wang et al., 2019).

5.2 Gaming

Games provide a unique sandbox to test the agentic behavior of large foundation models, pushing the boundaries of their collaborative and decision-making abilities. We describe three areas in particular that highlight agent's abilities to interact with human players and other agents, as well as their ability to take meaningful actions within an environment.

NPC Behavior. In modern gaming systems, the behavior of Non-Player Characters (NPCs) is predominantly dictated by predefined scripts crafted by developers. These scripts encompass a range of reactions and interactions based on various triggers or player actions within the gaming environment. In light of this situation, Agent AI is at the forefront of revolutionizing NPC technologies. By leveraging large foundation model, Agent AI can provide dynamic dialogues and refine behaviors based on player feedback and in-game data, significantly contributing to the evolution of NPC behavior in games.

Human-NPC Interaction. Agent AI plays a critical role in enhancing the interaction between human players and NPCs, offering a more immersive gaming experience. The conventional interaction paradigm is primarily one-dimensional, with NPCs reacting in a preset manner to player inputs. Agent AI, utilizing large foundation models, can analyze and learn from human behavior, providing more human-like interactions and increasing realism and immersion (Gong et al., 2023).

Agent-based Analysis of Gaming. Gaming is an integral part of daily life, estimated to engage half of the world's population (Intelligence, 2020) and exhibits a positive impact on mental health (Granic et al., 2014). Contemporary game systems, however, often exhibit deficiencies in interactions with human players due to primarily hand-crafted behaviors by game developers. In such a context, Agent AI proves valuable as a system that analyzes in-game text data, such as chat logs and player feedback, to identify patterns of player behavior and preferences, as well as analyzes image and video data from gaming sessions to understand user intent and actions.

Scene Synthesis for Gaming. Scene synthesis is essential for creating and enhancing immersive gaming environments, encompassing the generation of three-dimensional (3D) scenes, terrain creation, object placement, realistic lighting, and dynamic weather systems (Huang et al., 2023b). In modern games, providing vast open-world environments necessitates the use of procedural or AI-driven techniques for automated terrain generation. Agent AI, utilizing large foundation models, aids scene designers by formulating non-repeating, unique landscape design rules based on the designers' desires and the current scene, ensuring semantic consistency and variability of the generated assets. These models expedite object placement and assist in content generation, enhancing the design process.

5.3 Interactive Healthcare

In healthcare, Agent AI can help both patients and physicians by utilizing large foundation models in understanding the intent of the user, retrieving clinical knowledge, and grasping the undergoing human-to-human interaction, but not limited to these areas. Examples of application include:

Diagnostic Agents. LLMs as medical chatbots for patient diagnosis have gained attention for their potential to help triage and diagnose patients, providing equitable healthcare access to diverse populations (Lee et al., 2023). They offer a pathway to improve healthcare for millions, understanding various languages, cultures, and health conditions, with initial results showing promise using healthcare-knowledgeable LLMs trained on large-scale web data (Durante et al., 2024b,a). However, risks such as hallucination within medical contexts are notable challenges.

Knowledge Retrieval Agents. In the medical context, model hallucinations can be dangerous, potentially leading to serious patient harm or death. Approaches using agents for reliable knowledge retrieval (Peng et al., 2023) or retrieval-based text generation (Guu et al., 2020) are promising. Pairing diagnostic agents with medical knowledge retrieval agents can reduce hallucinations and improve response quality and preciseness.

Telemedicine and Remote Monitoring. Agent-based AI in Telemedicine and Remote Monitoring can enhance healthcare access, improve communication between healthcare providers and patients, and increase the efficiency of doctor-patient interactions (Amjad et al., 2023). Agents can assist in triaging messages from doctors, patients, and healthcare providers, highlighting important communications, and revolutionizing remote healthcare and digital health industries.

5.4 Interactive Multimodal Tasks

The integration of visual and linguistic understanding is a fundamental of Agent AI. Therefore, the development of Agent AI is closely linked to the performance of multimodal tasks, including image captioning, visual question answering, video language generation, and video understanding. Here are some tasks that have recently garnered significant interest:

Image and Language Understanding and Generation. Image-language understanding is a task that involves the interpretation of visual content in a given image with language and the generation of associated linguistic descriptions. This task is critical to the development of AI agents that can interact with the world in a more human-like manner. Some of most popular ones are image captioning (Lin et al., 2014; Sharma et al., 2018; Young et al., 2014; Krishna et al., 2016), referring expression (Yu et al., 2016; Karpathy et al., 2014), and visual question answering (Antol et al., 2015; Ren et al., 2015; Singh et al., 2019). This demands capabilities beyond object recognition, encompassing a deep understanding of spatial relationships, visual semantics, and integrating world knowledge for accurate descriptive and reasoning abilities.

Video-Language Understanding and Generation. Video captioning and storytelling involve generating coherent sentences for video frames, challenging due to the need for a comprehensive understanding of each frame and their interrelations. Recent advances leverage large foundation models for improved video-language generation, emphasizing the development of agent-aware text synthesis models for encoding sequences and generating cohesive paragraphs. Video understanding broadens image understanding to include dynamic content and requires agents to interact with visual, textual, and audio modalities. Key tasks include captioning, question answering, and activity recognition, focusing on temporal alignment, sequence handling, and complex activity interpretation. Agents also need to process audio cues like spoken words and background sounds to grasp a video's mood and nuances.

Parallel research explores generating scaled datasets from large models, then applying visual instruction tuning (Durante et al., 2024b,a; Li et al., 2023d; Zhu et al., 2023) on the generated data. Considerable audio, speech, and visual expert perception models are subsequently used to verbalize videos. Speech is transcribed with automatic speech recognition tools, and video descriptions and related data are produced with various tagging, grounding, and captioning models (Li et al., 2023e; Maaz et al., 2023; Chen et al., 2023; Wang et al., 2023c). These techniques demonstrate how instruction tuning video-language models on generated datasets may lead to enhanced video-reasoning and communication abilities.

Such agents would be able to understand the context of the video, identify the key steps, and generate a coherent summary of the procedure. This would not only enhance the interpretability of the model but also enable it to provide useful feedback or guidance to the user.

We expand upon more cross-modality and Mix-reality topic discussion in Appendix B.1, Appendix B.2 and B.3.

6 Deploying Agent AI

We believe that in order to develop a system that incorporates these elements, it is necessary to involve a wide range of experts and practitioners. For instance, there are the following important research areas:

Exploring new paradigms. The development of agent paradigms with integrated modalities (audio, image, text, sensor inputs) may address common issues in large-scale models, such as hallucinations and biases in their outputs, which will enhance their recognition and response capabilities for a wide variety of applications.

General-purpose end-to-end systems. Versatile and adaptable AI solutions can be driven by the development of end-to-end models that are trained with large-scale data.

Methodologies for grounding modalities. By integrating information across various modalities, we can enhance the coherence and efficacy of data processing. We expand on this topic in Appendix B.1.

Intuitive human interface. Developing intuitive human interfaces can facilitate effective and meaningful interactions between humans and agents.

Taming LLM/VLMs. Exploring new approaches can address common issues in large-scale foundation models, such as hallucinations and biases in their outputs.

Bridging the gap between simulation and real. The "sim-to-real" problem highlights the challenge of deploying AI agents trained in simulations to the real world, where discrepancies in conditions like disturbances and physical properties can degrade performance. To tackle these issues, strategies include:

- **Domain randomization** Introducing variability in the simulated environment to better prepare the model for real-world unpredictability (Tobin et al., 2017; Saito et al., 2022).
- **Domain adaptation** Bridging sim-to-real gap by training on both simulated and real-world data (Zhu et al., 2017a; Rao et al., 2020; Ho et al., 2021).
- **Improvement of simulation** Enhancing simulation fidelity through better replication of real-world conditions (Zhu et al., 2017b; Allevato et al., 2020; Martinez-Gonzalez et al., 2020; Müller et al., 2018; Shah et al., 2018; Sasabuchi et al., 2023).

Multi-Agent. Agent AI interaction is currently still a complex process that requires a combination of multiple skills. The current human-machine interaction systems inside multi-agents are primarily effectiveness of cooperation rule-based. They do have intelligent behaviors in response to human/user actions and possess web knowledge to some extent (Gong et al., 2023). The kind multi agents interactions are very important in the agent development to enable specific behaviors in the agent system design.

Agent Infrastructure and System. Agent-based AI is a large and fast-growing community within the domains of entertainment, research, and industry. The development of large foundation models has significantly improved the performance of agent AI systems. However, creating agents in this vein is limited by the increasing effort necessary to create high-quality datasets and overall cost. In industry, building high-quality agent infrastructure has significantly impacted multi-modal agent copilots by using advanced hardware, diverse data sources, and powerful software libraries (Gong et al., 2023). The rising prevalence of Agent AI underscores the need for robust infrastructure to facilitate their training, evaluation, and deployment. In response to this need, we are introducing a dedicated track for agent research focusing on the infrastructure and methodologies pertinent to the development, evaluation, and deployment of Agent AI. We expect this track will attract a significant number of submissions centered on the efficiency and optimization of agent systems. Agent AI infrastructure is intended to ensure that the broader community can readily access and benefit from these contributions, thereby fostering further advancements in the field.

We expand on biases and hallucinations in Appendix C and D respectively.

7 Challenges for Agent AI

In this paper, we put special emphasis on discovering the current agent AI limitation, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction.

Achievement of the Agent AI still have some challenges, especially considering the dynamic system with high modality observations in the physical world. There still exist a number of challenges that need to be addressed, including but not limited to: 1) unstructured environments, where current visual inputs affect both high-level intents and low-level actions of the embodied agent given the same goal instruction; 2) empathy for agent, when open sets of objects, which require the agent's decision-making module to use common sense knowledge that is hard to encode manually; 3) multi-agent interactions and collaborations, which require the agent to understand and operate on more than just template-based commands, but also a context of goals, constraints, and partial plans expressed in everyday language. To enable a more comprehensive approach to these complex challenges, the inclusion of researchers and practitioners from a broader range of fields is critical. 4) Emergent ability for embodied large agent foundation model. We aspire to broaden our collective understanding of the potential and limitations of Agent Paradigm by leveraging our unique and diverse perspectives. We strongly believe that this proposed new agent paradigm will not only enrich the perspectives of individual practitioners, but will also enhance the community's collective knowledge and promote a holistic view that is more inclusive of the wide-ranging challenges faced by future agent AI.

8 Emergent Abilities

Despite the growing adoption of interactive agent AI systems, the majority of proposed methods still face a challenge in terms of their generalization performance in unseen environments or scenarios. Current modeling practices require developers to prepare large datasets for each domain to finetune/pretrain models; however, this process is costly and even impossible if the domain is new. To address this issue, we propose building interactive agents that leverage the knowledge-memory of general-purpose foundation models (ChatGPT, Dall-E, GPT-4, etc.) for novel scenarios, specifically for generating a collaborative space between humans and agents. We discover an emergent mechanism— which we name Mixed Reality with Knowledge Inference Interaction—that facilitates collaboration with humans to solve challenging tasks in complex real-world environments and enables the exploration of unseen environments for adaptation to virtual reality. For this mechanism, the agent learns i) micro-reactions in cross-modality: collecting relevant individual knowledge for each interaction task (e.g., understanding unseen scenes) from the explicit web source and by implicitly inferring from the output of pretrained models; ii) macro-behavior in reality-agnostic: improving interactive dimensions and patterns in language and multi-modality domains, and make changes based on characterized roles, certain target variable, influenced diversification of collaborative information in mixed-reality and LLMs. We investigate the task of knowledge-guided interactive synergistic effects to collaborated scene generation with combining various OpenAI models, and show promising results of how the interactive agent system can further boost the large foundation models in our setting. It integrates and improves the depth of generalization, conscious and interpretability of a complex adaptive AI systems.

9 Impact Statement

Agent AI paradigm is to create general-purpose agents that can work alongside humans in both real and virtual environments. This paradigm therefore intends to have a very broad impact, possibly affecting all members of society. Agent AI framework emphasizes the integration of agents into the wider environment across a variety of settings, such as gaming, robotics, healthcare, and long-video understanding. Specifically, the development of multimodal agents in gaming could lead to more immersive and personalized gaming experiences, thereby transforming the gaming industry. In robotics, the development of adaptive systems could revolutionize industries ranging from manufacturing to agriculture, potentially addressing labor shortages and improving efficiency. In healthcare, the use of large foundation model as diagnostic agents or patient care assistants could lead to more accurate diagnoses, improved patient care, and increased accessibility to medical services, particularly in underserved areas. Furthermore, the ability of these models to interpret long-form videos could have far-reaching applications, from enhancing online learning to improving technical support services. In general, the Agent AI framework will have significant downstream effects on a wide range of industries and people across the world.

We must also highlight the diverse and complex challenges that come with implementing AI agents across a wide variety of environments and situations. For instance, there are many limitations and potential hazards linked to Agentic AI systems when they are developed for specialized sectors such as healthcare diagnostics. In this domain, issues like dangerous hallucinations in AI behavior can pose significant risks, highlighting the critical need for meticulous design and testing. However, these specific challenges may not be equally relevant or noticeable when considering AI agents crafted for the gaming industry. In such recreational fields, developers might instead prioritize tackling different hurdles, such as the need for AI to perform more open-ended generation and exhibit creativity, adapting dynamically to unpredictable gameplay scenarios and player interactions.

10 Conclusion

Our proposed Agent AI focuses on advanced multimodal systems that interact effectively within both physical and virtual environments and facilitate effective interaction with humans. This paper aims to unite researchers to deepen the discourse on Agent AI, cutting across various AI disciplines including agent paradigms, foundation models, infrastructures, and systems. Our goal is to enrich the scientific comprehension of Agent AI and explore the potential of embodied agents within the realm of holistic intelligence research. This endeavor positions us to leverage emerging foundational models effectively.

Ethical Consideration

Agent AI systems have many applications. In addition to interactive AI, grounded multimodal models could help in generating training datasets for robots and AI agents, and assist in productivity applications, helping to re-play or paraphrase scenario, predict actions in novel scenarios, or synthesize 3D or 2D scenes. Fundamental advances in agent AI help contribute towards these goals and many would benefit from a greater understanding of how to model embodied and empathetic behavior in a simulated environment or the real world. Therefore, there are many applications that have positive benefits.

However, this technology could also be used by bad actors. Agent AI systems that generate content can be used to manipulate or deceive people. Therefore, it is very important that this technology is developed in accordance with responsible AI guidelines. For example, explicitly communicating to users that content is generated by an AI system and providing the user with controls in order to customize such a system. It is possible the Agent AI could be used to develop new methods to detect manipulative content - partly because it is rich with hallucinations that emerge from large foundation models - and thus help address another real world problem.

For example, ethical deployment of large agents foundation models, especially in sensitive domains like healthcare, is paramount. AI agents trained on biased data could potentially worsen health disparities by providing inaccurate diagnoses for underrepresented groups. Moreover, the handling of sensitive patient data by AI agents raises significant privacy and confidentiality concerns. In the gaming industry, AI agents could transform the role of developers, shifting their focus from scripting non-player characters to refining agent learning processes. Similarly, adaptive robotic systems could redefine manufacturing roles, necessitating new skill sets rather than replacing human workers. Navigating these transitions responsibly is vital to minimize potential socio-economic disruptions.

Furthermore, the agent AI focuses on learning collaborative policies in simulation and there is some risk of directly applying the policy to the real world due to the distribution shift. Robust testing and continuous safety monitoring mechanisms should be put in place to minimize risks of unpredictable behaviors in real-world scenarios.

Limitations

The main thesis of our work is that the Agent AI formulation helps to bring the field of AI back to its roots in holistic intelligence. However, there are still many unknowns within the Agent AI paradigm. Existing foundation models exhibit biases and hallucinations, and it is unclear whether these can be resolved through scaling up model and dataset sizes or if these are fundamental limitations of Agent AI.

We also acknowledge that there are many additional challenges in this field that we have not covered in Section 7. As a growing field with a potential for major impact, we believe that the development of Agent AI must include a diverse range of perspectives across disciplines to ensure that it has a positive impact on humanity.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022a. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022b. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Adam Allevato, Elaine Schaertl Short, Mitch Pryor, and Andrea Thomaz. 2020. Tunenet: One-shot residual tuning for system identification and sim-to-real robot task transfer. In *Conference on Robot Learning*, pages 445–455. PMLR.

Ayesha Amjad, Piotr Kordel, and Gabriela Fernandes. 2023. A review on innovation in healthcare sector (telehealth) through artificial intelligence. *Sustainability*, 15(8):6655.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. 2023. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. 2023. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2023. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. *arXiv preprint arXiv:2309.10309*.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020a. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. 2020b. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. 2023. Videollm: Modeling video sequence with large language models.

Kezhen Chen, Qiuyuan Huang, Daniel McDuff, Xiang Gao, Hamid Palangi, Jianfeng Wang, Kenneth Forbus, and Jianfeng Gao. 2021. Nice: Neural image commenting with empathy. In *EMNLP 2021*.

Kezhen Chen, Qiuyuan Huang, Hamid Palangi, Paul Smolensky, Kenneth D. Forbus, and Jianfeng Gao. 2020. Mapping natural-language problems to formal-language solutions using structured neural representations. In *ICML 2020*.

Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Foerster. 2023. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning Representations*.

Gautier Dagan, Frank Keller, and Alex Lascarides. 2023. Dynamic planning with a llm. *arXiv preprint arXiv:2308.06391*.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. 2020. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174.

Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. 2023. Can an embodied agent find your" cat-shaped mug"? llm-based zero-shot object navigation. *arXiv preprint arXiv:2303.03480*.

Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. 2022. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. 2023. Video language planning. *arXiv preprint arXiv:2310.10625*.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024a. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, Demetri Terzopoulos, Ade Famoti, Noboru Kuno, Ashley J. Llorens, Hoi Vo, Katsushi Ikeuchi, Fei-Fei Li, Jianfeng Gao, Naoki Wake, and Qiuyuan Huang. 2024b. Agent foundation model.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2021. Manipulathor: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4497–4506.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181.

Jianfeng Gao, Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Heung-Yeung Shum. 2020. Robust conversational ai with grounded text generation. *arXiv preprint arXiv:2009.03457*.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. 2021. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4:265–293.

Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. 2023. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. 2023. Rvt: Robotic view transformer for 3d object manipulation. *arXiv preprint arXiv:2306.14896*.

Isabela Granic, Adam Lobel, and Rutger CME Engels. 2014. The benefits of playing video games. *American psychologist*, 69(1):66.

Liangke Gui, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022a. Vlc: Training vision-language transformers from captions.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022b. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL 2022. Long paper, Oral.*

Rodrigo Longhi Guimarães, André Schneider de Oliveira, João Alberto Fabro, Thiago Becker, and Vinícius Amilgar Brenner. 2016. Ros navigation: Concepts and tutorial. *Robot Operating System (ROS) The Complete Reference (Volume 1)*, pages 121–160.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Huy Ha, Pete Florence, and Shuran Song. 2023. Scaling up and distilling down: Language-guided robot skill acquisition. *arXiv preprint arXiv:2307.14535*.

Laura Birka Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. 2023. Large language models in textual analysis for gesture selection. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, pages 378–387.

Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. 2021. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926. IEEE.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. 2023a. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE.

Qiuyuan Huang, Jae Sung Park, Abhinav Gupta, Paul Bennett, Ran Gong, Subhojit Som, Baolin Peng, Owais Khan Mohammed, Chris Pal, Yejin Choi, et al. 2023b. Ark: Augmented reality with knowledge interactive emergent ability. *arXiv preprint arXiv:2305.00970*.

Qiuyuan Huang, Pengchuan Zhang, Oliver Wu, and Lei Zhang. 2018. Turbo learning for captionbot and drawingbot. In *NeurIPS 2018*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022b. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*.

DFC Intelligence. 2020. Global video game audience reaches 3.7 billion. `https://www.dfcint.com/global-video-game-audience-reaches-3-7-billion/`. Accessed: 2024-02-05.

Stephen James and Andrew J Davison. 2022. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):1612–1619.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2022. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. *arXiv*.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27.

Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. 2023. Gen2sim: Scaling up robot learning in simulation with generative models. *arXiv preprint arXiv:2310.18308*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*.

K Niranjan Kumar, Irfan Essa, and Sehoon Ha. 2023. Words into action: Learning diverse humanoid robot behaviors using language guided iterative motion refinement. *arXiv preprint arXiv:2310.06226*.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. 2023a. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Jiachen Li, Qiaozi Gao, Michael Johnston, Xiaofeng Gao, Xuehai He, Suhaila Shakiah, Hangjie Shi, Reza Ghanadan, and William Yang Wang. 2023c. Mastering robot manipulation with multimodal prompts through pretraining and multi-task fine-tuning. *arXiv preprint arXiv:2310.09676*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

KunChang Li, Yinan He, Wang Yi, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023e. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2023f. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. In *arXiv preprint arXiv:2209.07753*.

Xiwen Liang, Liang Ma, Shanshan Guo, Jianhua Han, Hang Xu, Shikui Ma, and Xiaodan Liang. 2023. Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation. *arXiv preprint arXiv:2306.10322*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. *Proceedings of ECCV*.

C Karen Liu and Dan Negrut. 2021. The role of physics-based simulators in robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:35–58.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.

Pablo Martinez-Gonzalez, Sergiu Oprea, Alberto Garcia-Garcia, Alvaro Jover-Alvarez, Sergio Orts-Escolano, and Jose Garcia-Rodriguez. 2020. Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *Virtual Reality*, 24:271–288.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334.

Meta Fundamental AI Research, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, et al. 2023. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Matthias Müller, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem. 2018. Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision*, 126:902–919.

Zhe Ni, Xiao-Xin Deng, Cong Tai, Xin-Yue Zhu, Xiang Wu, Yong-Jin Liu, and Long Zeng. 2023. Grid: Scene-graph-based instruction-driven robotic task planning. *arXiv preprint arXiv:2309.07726*.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Meenal Parakh, Alisha Fong, Anthony Simeonov, Abhishek Gupta, Tao Chen, and Pulkit Agrawal. 2023. Human-assisted continual robot learning with foundation models. *arXiv preprint arXiv:2309.14321*.

Jae Sung Park, Jack Hessel, Khyathi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, and Yejin Choi. 2023a. Multimodal agent – localized symbolic knowledge distillation for visual commonsense models. In *NeurIPS 2023*.

Jae Sung Park, Jack Hessel, Khyathi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Youngjae Yu, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, and Yejin Choi. 2023b. Localized symbolic knowledge distillation for visual commonsense models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023c. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8494–8502.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. 2022. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900.

Shreyas Sundara Raman, Vanya Cohen, David Paulius, Ifrah Idrees, Eric Rosen, Ray Mooney, and Stefanie Tellex. 2023. Cape: Corrective actions from precondition errors using large language models. In *2nd Workshop on Language and Robot Learning: Language as Grounding*.

Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. 2020. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11157–11166.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Antoni Rosinol, John J Leonard, and Luca Carlone. 2022. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*.

Daichi Saito, Kazuhiro Sasabuchi, Naoki Wake, Atsushi Kanehira, Jun Takamatsu, Hideki Koike, and Katsushi Ikeuchi. 2023. Constraint-aware policy for compliant manipulation.

Daichi Saito, Kazuhiro Sasabuchi, Naoki Wake, Jun Takamatsu, Hideki Koike, and Katsushi Ikeuchi. 2022. Task-grasping from a demonstrated human strategy. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 880–887.

Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. 2023. Diverse conventions for human-AI collaboration. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kazuhiro Sasabuchi, Daichi Saito, Atsushi Kanehira, Naoki Wake, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Task-sequencing simulator: Integrated machine learning to execution simulation for robot manipulation. *arXiv preprint arXiv:2301.01382*.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Brennan Shacklett, Luc Guy Rosenzweig, Zhiqiang Xie, Bidipta Sarkar, Andrew Szot, Erik Wijmans, Vladlen Koltun, Dhruv Batra, and Kayvon Fatahalian. 2023. An extensible, data-oriented architecture for high-performance, many-world simulation. *ACM Trans. Graph.*, 42(4).

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023a. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR.

Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. 2023b. Mutex: Learning unified policies from multimodal task specifications. *arXiv preprint arXiv:2309.14320*.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2023. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Sumedh A Sontakke, Jesse Zhang, Sébastien MR Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. 2023. Roboclip: One demonstration is enough to learn robot policies. *arXiv preprint arXiv:2310.07899*.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pages 477–490. PMLR.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hitoshi Teshima, Naoki Wake, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. 2022. Deep gesture generation for social robots using type-specific libraries. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8286–8291. IEEE.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. 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)*, pages 23–30. IEEE.

Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W Gupta, Mubbasir Kapadia, and Marynel Vázquez. 2022. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4):11047–11054.

Naoki Wake, Riku Arakawa, Iori Yanokura, Takuya Kiyokawa, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2021. A learning-from-observation framework: One-shot robot teaching for grasp-manipulation-release household operations. In *2021 IEEE/SICE International Symposium on System Integration (SII)*. IEEE.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023a. Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.11753*.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023b. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*, 11:95060–95078.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023c. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023d. Gpt models meet robotic applications: Co-speech gesturing chat system. *arXiv preprint arXiv:2306.01741*.

Borui Wang, Qiuyuan Huang, Budhaditya Deb, Aaron L. Halfaker, Liqun Shao, Daniel McDuff, Ahmed Awadallah, Dragomir Radev, and Jianfeng Gao. 2023a. Logical transformers: Infusing logical structures into pre-trained language models. In *Proceedings of ACL 2023*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023b. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Weng, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR 2019*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Internvid: A large-scale video-text dataset for multimodal understanding and generation.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. 2023d. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023e. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

P. H. Winston. 1972. The m.i.t. robot. In D. Michie, editor, *Machine Intelligence 7*. Edinburgh University Press, Edinburgh, Scotland.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. 2023a. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023b. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Yang Ye, Hengxu You, and Jing Du. 2023. Improved trust in human-robot collaboration with chatgpt. *IEEE Access*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. 2023. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. 2021. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR.

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. 2021. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15130–15140.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Zhide Zhong, Jiakai Cao, Songen Gu, Sirui Xie, Weibo Gao, Liyi Luo, Zike Yan, Hao Zhao, and Guyue Zhou. 2023. Assist: Interactive scene nodes for scalable and realistic indoor simulation. *arXiv preprint arXiv:2311.06211*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2023a. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*.

Haoyu Zhou, Mingyu Ding, Weikun Peng, Masayoshi Tomizuka, Lin Shao, and Chuang Gan. 2023b. Generalizable long-horizon manipulations with large language models. *arXiv preprint arXiv:2310.02264*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023c. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Shaojun Zhu, Andrew Kimmel, Kostas E Bekris, and Abdeslam Boularias. 2017b. Fast model identification via physics engines for data-efficient policy search. *arXiv preprint arXiv:1710.08893*.

Appendices for
Agent AI Towards a Holistic Intelligence

A Intention Information and Manipulation for Embodied Action

Language Conditioned internet action instruction entails the ability of a robotic system to interpret and execute tasks based on language instructions. This aspect is particularly crucial for creating intuitive and user-friendly interfaces for human-robot interaction. Through natural language commands, users can specify goals and tasks to robots in a manner similar to human-human communication (Wang et al., 2019), thereby lowering the barrier to operating robotic systems. In a practical scenario, for instance, a user could instruct a service robot to "pick up the red apple from the table," and the robot would parse this instruction, identify the referred object and execute the task of picking it up (Wake et al., 2023b). The core challenge lies in developing robust natural language processing and understanding algorithms that can accurately interpret a wide array of instructions, ranging from direct commands to more abstract directives, and enable the robot to convert these instructions into actionable tasks. Furthermore, ensuring that robots can generalize these instructions across diverse tasks and environments is critical for enhancing their versatility and utility in real-world applications. The use of language input to guide robot's task planning has gained attention in the context of a robot framework called Task and Motion Planning (Garrett et al., 2021).

In addition, (Durante et al., 2024b)learn about the intricate challenges of large action models for embodied systems e.g., robotic. It begin with a low-level action manipulation foundational models, it explore solutions to issues such as action resignation, adaptability to dynamic environments, and the efficient management of high-dimensional action spaces. When transfer to next phase, we implement and refine algorithms, ensuring scalability and effectiveness in simulations on our server. Develop and implement foundational algorithms for large action models, emphasizing efficiency and scalability. Focus on addressing issues related to pre-training, fine-tuning, and model optimization. Conduct initial simulations on Azure to validate algorithmic concepts. The ultimate objective in the third phase, is to optimize and validate these algorithms in real-world scenarios, exploring diverse applications and contributing to the evolution of large action models for General Purpose Robotics. Optimize and validate the algorithms in real-world scenarios with large action models on Phoenix. Explore applications in diverse domains, ensuring robustness and scalability. Refine algorithms based on real-world evaluation feedback and scale for broader cloud deployment in the embodied system.

B Agent for Cross-modality and Mix-reality

B.1 Agents for Cross-modal Understanding

Multi-modal understanding is a significant challenge for creating generalist AI agents due to the lack of large-scale datasets that contain vision, language, and agent behavior. More generally, training data for AI agents is often modality specific. This results in most modern multi-modal systems using a combination of frozen submodules. Some notable examples are Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023d), VLC (Gui et al., 2022a) and ArK (Huang et al., 2023b), all of which utilize a frozen LLM and frozen visual encoder. These submodules are trained individually on separate datasets, and then adaptation layers are trained to encode the visual encoder into the LLM embedding space. In order to make further progress for cross-modal understanding for AI agents, it is likely that the strategy of using frozen LLMs and visual encoders will need to change. Indeed, RT-2, a recent visual-language model that is capable of taking actions within the domain of robotics showed significantly improved performance when jointly tuning the visual encoder and LLM for robotics and visual-language tasks (Brohan et al., 2023).

B.2 Agents for Cross-domain Understanding

A key challenge for creating generalist agents is the distinctive visual appearance and disparate action spaces across different domains. Humans possess the capability to interpret images and videos from various sources, including the real world, video games, and specialized domains such as robotics and healthcare (Durante et al., 2024a), once they become familiar with the specific details of these areas. However, existing LLMs and VLMs often demonstrate significant differences between the data they were trained on and the varied domains in which they are applied. And notably, training agent models to predict specific actions presents a considerable challenge when trying to develop a single policy that can effectively learn multiple control systems across domains (Huang et al., 2023b). Generally, the approach most modern works take when applying systems within specific domains is to start from a pretrained foundation model and then finetune a separate model for each specific domain. This fails to capture any commonalities between domains and results in a smaller total set of data used for training instead of leveraging each domain's data.

B.3 Interactive agent for cross-modality and cross-reality

Developing AI agents that can successfully understand and perform tasks across different realities is an on-going challenge that has seen some recent success for image and scene generation (Huang et al., 2023b). In particular, it is challenging for agents to simultaneously understand real-world and virtual reality environments due to their visual dissimilarities and separate environment physics. Within the context of cross-reality, Sim to Real transfer is a particularly important problem when using simulation-trained policies for real-world data, which we discuss in the next section.

C Bias

AI agents based on LLMs or LMMs (large multimodal models) have biases due to several factors inherent in their design and training process. When designing these AI agents, we must be mindful of being inclusive and aware of the needs of all end users and stakeholders. In the context of AI agents, *inclusivity* refers to the measures and principles employed to ensure that the agent's responses and interactions are inclusive, respectful, and sensitive to a wide range of users from diverse backgrounds. Despite these measures, AI agents still exhibit biases. Ongoing efforts in agent AI research and development are focused on further reducing these biases and enhancing the inclusivity and fairness of agent AI systems. Despite these measures, AI agents still exhibit biases. Ongoing efforts in agent AI research and development are focused on further reducing these biases and enhancing the inclusivity and fairness of agent AI systems. Despite these efforts, it's important to be aware of the potential for biases in responses and to interpret them with critical thinking. Continuous improvements in AI agent technology and ethical practices aim to reduce these biases over time. One of the overarching goals for inclusivity in agent AI is to create an agent that is respectful and accessible to all users, regardless of their background or identity.

D Hallucinations

Agents that generate text are often prone to hallucinations, which are instances where the generated text is nonsensical or unfaithful to the provided source content (Raunak et al., 2021; Maynez et al., 2020). Hallucinations can be split into two categories, *intrinsic* and *extrinsic* (Ji et al., 2023). Intrinsic hallucinations are hallucinations that are contradictory to the source material, whereas extrinsic hallucinations are when the generated text contains additional information that was not originally included in the source material.

Some promising routes for reducing the rate of hallucination in language generation involve using retrieval-augmented generation (Lewis et al., 2020; Shuster et al., 2021) or other methods for grounding natural language outputs via external knowledge retrieval (Dziri et al., 2021; Peng et al., 2023). Generally, these methods seek to augment language generation by retrieving additional source material and by providing mechanisms to check for contradictions between the generated response and the source material.

Within the context of multi-modal agent systems, have multimodality been shown to hallucinate as well (Zhou et al., 2023c). One common cause of hallucination for vision-based language-generation is due to the over-reliance on co-occurrence of objects and visual cues in the training data (Rohrbach et al., 2018). AI agents that exclusively rely upon pretrained large foundation models and use limited environment-specific finetuning can be particularly vulnerable to hallucinations since they rely upon the internal knowledge-base of the pretrained models for generating actions and may not accurately understand the dynamics of the world state in which they are deployed.