Position Paper: Agent AI Towards a Holistic Intelligence

Qiuyuan Huang [*1 ▶] , **Naoki Wake** [*1 ▶ ◊] , **Bidipta Sarkar** [2 1 †] , **Zane Durante** [2 1 †] , **Ran Gong** [1 3 †] , **Rohan Taori** [2 1 †] , **Yusuke Noda** [1] , **Demetri Terzopoulos** [3] , **Noboru Kuno** [1] , **Ashley Llorens** [1] , **Hoi Vo** [1 §] , **Katsu Ikeuchi** [1 §] , **Li Fei-Fei** [2 §] , **Jianfeng Gao** [1 §]

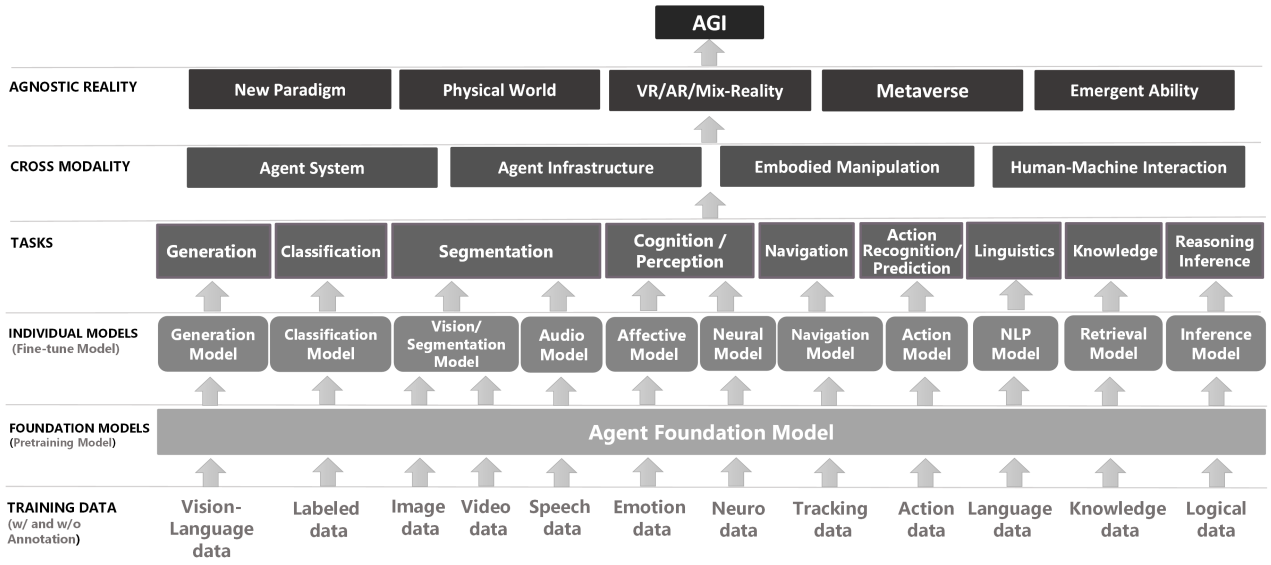

Figure 1. Overview of an Agent AI system. It is applicable to multiple domains and provides an agent foundation model for interactive manipulation and embodied operations. Agent AI functions in both physical and virtual worlds by training on cross-modal data that is obtained through interactions between diverse environments. Agent AI offers a promising approach to unify a broad range of applications and capabilities for infrastructure and system, and it is emerging as a promising avenue a route towards Artificial General Intelligence (AGI) using a new agent AI paradigm.

Abstract

Recent advancements in large foundational models have remarkably enhanced our understanding of sensory information in open-world environments. At this pivotal moment, it is crucial to the AI research trend toward excessive reductionism and returning to the AI principles inspired by the holistic philosophy of Aristotle. 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 diagnostic systems. We emphasize the importance of integrating recent large foundational models to enhance intelligence and interaction capabilities. Furthermore, we discuss how agents exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. This paper we aim to broaden the research community's perspective on achieving holistic intelligence, while highlighting the need for an integrated approach that considers the agent's purpose, functionality, and interaction. Finally, we reflect on a deeper discussion of these Agent AI topics from a mainstream and interdisciplinary perspective. This discussion illustrates AI cognition and consciousness within the scope of scientific discourse, and may serves as a basis for future research directions and social influences.

[*]Equal Contribution. [▶]Project Lead. [§] Equal Advisor. [◊]Corresponding Author. [1]Microsoft Research, Redmond; [2]Stanford University; [3]University of California, Los Angeles. [†]Work done while interning or researching part-time at Microsoft Research, Redmond.

1. Introduction

Historically, the AI systems were defined at the Dartmouth Conference as artificial life forms that could collect information from the environment and interact with the environment. For example, MIT's Minsky group, inspired by this definition, built a system, called "copy demo", which observed a block world and successfully completed the same structure; the system consists of observation and interaction modules. These early studies revealed that each module itself was quite challenging and further research were necessary. As the results, each module, based on the divide-and-conquer approach coming from Rene Descartes Reductionism, become specialized and fragmented to conduct research. This too-much-reductionism approach provides the situation that Cambrian explosion of each field occurs and the overall goals of the AI research became less clear.

To address this trend, it is necessary to return the fundamentals based on Aristotelian Holism. Fortunately, leveraging recent progress of Large Language Models (LLMs) and Visual Language Models (VLMs) has made it possible to reconstruct such AI agent along the Holismic idea. Seizing this opportunity, our proposed framework, "Agent AI", emphasize to develop a comprehensive intelligence system that integrates language proficiency, visual cognition, context memory, intuitive reasoning and adaptability. It explore the potential completing this synthesis using LLM and VLM. During this exploration, consideration is also given to revisiting the system's design based on Aristotle's final cause (why the system exists), which may have been overlooked in the initial round of AI agent development in 70s. Specifically, we define Agent AI as "an intelligent agent capable of autonomously executing appropriate and seamless actions based on sensory input, whether in a physical, virtual, or mixed-reality environment."

Importantly, an embodied agent is conceptualized as a collaborative system, where it communicates with humans or environments with its perception capabilities and employ a set of vast actions based on human needs. This is the reason why we consider that the advance of LLMs and VLMs (OpenAI, 2023) will make a significant contribution to Agent AI, enabling systems to parse and infer human intent from natural-form instructions and images.

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). In this paper, we introduce the cognitive aspects for Agent AI, and review recent literature in Agent AI domains including robotics, gaming, and healthcare. This approach allow us to illustrate how the development of those technologies is bringing the agent closer to holistic ideal. Furthermore, we 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.

2. Agent AI Paradigm

Agent AI new paradigm represents a change in thinking in embodied intelligence, emphasizing the importance of complex dynamics, and an integrated approach to interactive intelligence. This approach is motivated by the belief that true intelligence arises from the intricate interplay between learning, memory, action, observation, planning, perception, and cognition in a interactive decision with consciousness.

As shown in the Fig. 2, a new Agent AI paradigm aims to explore the complex challenges towards general purpose agent and interactive intelligence by leveraging interdisciplinary in the computer science, biological physics, cognition science, medical health, and moral philosophy. In this paper, we argue that consciousness/cognition of AI is best assessed by drawing on neuroscientific theories of consciousness. We describe prominent theories of this kind and investigate their implications for agent AI.

We define the Agent AI as *"any intelligent agent capable of autonomously taking suitable and seamless action based on sensory input, whether in the physical world or in a virtual or mixed-reality environment representing the physical world."*. Importantly, an embodied agent is conceptualized as a collaborative system, where it communicates with humans with its vision-language capabilities and employ a set of vast actions based on human needs. In this manner, embodied agents are expected to mitigate cumbersome tasks in virtual reality and physical world.

Despite numerous gaps between current technologies and holistic intelligence, the recent advancements in LLMs/VLMs have brought society closer to the idea that such a system is within reach. What steps are required to achieve this ultimate goal? In light of traditional AI philosophy, we believe that successful Agent AI systems require several key components:

Cognitive Aspects. The concept of holistic intelligence focuses not only on the exceptional performance of individual components (e.g., image recognition, language processing, task planning) but also on the utility of the system as a whole. Consider a scenario where a robot, right after being unboxed, begins to communicate instantly 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 with only a single component. Moreover, cog-

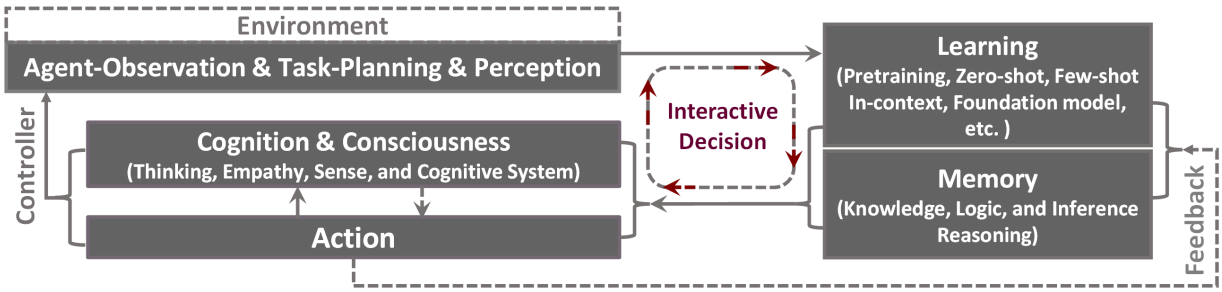

Figure 2. An Agent AI paradigm for supporting embodied multi-modal generalist agent systems. There are 5 main modules as shown: (1) Agent in Environment and Perception with task-planning and observation, (2) Agent learning, (3) Memory, (4) Action, and (5) Cognition and Consciousness (we use "consciousness" to imply a degree of awareness of an agent's state and surroundings). A key difference between our approach and some previous interactive strategies is that, after training, the agent's action will directly impact task planning, as the agent does not need to receive feedback from the environment to plan its next actions.

nitive functions, such as high-level task planning, human communication, a deep understanding of the relationships between the environment and actions, and their integration are necessary.

We can build a neuro-cognitive module that can be deployed onto its embodied robots, and generalizable into other agents through a cloud service. The deployed cognitive on the infinite agent will give the ability to understand and respond to dynamic, real-world situations, making them potentially more versatile and adaptive in complex environments.

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. 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.

Interaction. In general, real-world operations cannot be completed in one shot and thus require multi-round interactions between humans or the environment and the agent. Enabling fluent interactions is key to effective operation.

Memory. Long-term memory enables the Agent to remember specific operations adaptable to the environment or user preference. In contrast, short-term memory refers to the history of actions and perception results during an operation. Short-term memory enables the system to replan and consider next-step actions based on history.

Learning. An intelligent agent can adapt to a new environment by acquiring new knowledge and updating its skills. To this end, the agent should learn from human demonstrations. Additionally, the agent should always be under human supervision for safety. In case it encounters a difficult situation, it should ask the user for help and further instructions.

Achieving embodied agents that incorporate these elements is not straightforward. In the section 4, we will introduce a specific example that embodies these aspects. In Section 6, we will discuss the main challenges and necessary actions, including ethical concerns in Agent AI research.

3. Agent AI Foundation Model Mechanism

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

3.1. Agent Transformer

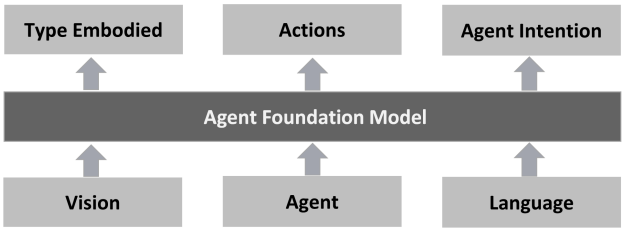

Figure 3. Overview of an interactive agent foundation model framework. The transformer is designed to process multi-modal information that conveys various levels of abstraction. This approach facilitates a comprehensive understanding of the context, thus enhancing coherent actions. Through learning across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

We have designed a multimodal encoding transformer (Fig. 3) as the model that allows interactive agent actions based on multimodal information. This model is initialized with two pre-trained submodules, namely, CLIP ViT-B16, which is used to initialize our visual encoder, and OPT-125M, which is used to initialize our action and language model.

In our approach, each frame in a video is encoded as visual features. To facilitate cross-modal information sharing, we train an additional linear layer that transforms the embeddings of our visual encoder into the token embedding space of our transformer model. This allows us to make predictions for text or action tokens based on both a text prompt and a single video frame.

We also incorporate historical data into our model by including previous actions and visual frames as input during pre-training. As a result, for any given time step, we can predict the action token based on the text prompt, the visual frames up to that point, and the actions taken up to that point. This approach allows our 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

Agent with Reinforcement Learning (RL). To facilitate human-AI interaction, we can develop a embodied agent which uses an emerging mechanism for generating and understanding scenes in virtual or real worlds with the reinforcement learning module stores human-AI interactions from which the human intent feedback. The embodied agent is trained via reinforcement learning to incorporate feedback using the reward like GPT-X setting. We can use the actor-critic algorithm PPO (Schulman et al., 2017) to update the parameters of the agent using its own version.

Generation Agent with Imitation Learning (IL). Generally speaking, a trained interactive agent can be used to perform simulation/VR scene generation. Note the agent requires an multi-model information (image/video and language) to generate relevant memory for the interaction model. For example, we can use a text-to-image generative model, GPT-4V, to reconstruct the physical anchor view which is further used to extract the desired memory. LLM/VLM implicitly serves as the vision-memory source that contains the visual prior memory of what we can imagine from the task planing. The embodied agent then takes as input the original language information and the generated simulation information to retrieve memory and outputs a action prediction tuple, while foundation model can generates new memory-enhanced prompt using the agent output.

To generate the simulation transfer physical world scene from memory/task planing prompt, we use can use the LLM/VLM to output user intent instruction that is then rendered using a similation rendering engine. We use the prompt and action instruction in foundation model to generate the spatial arrangement in the simulation environment. We perform experiments with foundation model as the low-level action generation model, and we can use the foundation model to load the simulation models viewable in the simulation environment. More information about generating the prompt to run the simulation robot can be referenced in (Wake et al., 2023c) and (Wake et al., 2023b).

4. Agent AI Categorization

Agent AI refers to integrated AI systems that LLMs/VLMs. Consequently, many of the AI systems based on VLMs or LLMs proposed in recent years can be categorized and associated with Agent AI subcatogiries. This section reviews recent related research, elucidating the Agent AI aspects they capture. We emphasize the significance of integrating VLMs/LLMs and clarify the research domains encompassed by Agent AI.

4.1. Embodied Agents

Embodied AI refers to agent systems that particularly emphasize interaction with the environment. This field is constantly evolving due to the development of perception models such as image recognition, speech recognition, and natural language processing, along with the advancement of reinforcement learning techniques.

1. **Action Agents:** Agents performing physical actions in simulated or real-world environments, divided into gaming AI and robotics. (Meta Fundamental AI Research (FAIR) Diplomacy Team et al., 2022; Park et al., 2023b; Huang et al., 2022a; Wang et al., 2023b; Yao et al., 2023; Li et al., 2023c; Ahn et al., 2022a; Huang et al., 2022b; Liang et al., 2022; Wang et al., 2023e; Baker et al., 2022; Driess et al., 2023; Brohan et al., 2023)

2. **Interactive Agents:** A broader category than action agents, these agents interact with the world through various means, not limited to physical actions, and include applications in diagnostics and knowledge retrieval. (Lee et al., 2023; Peng et al., 2023)

4.2. Simulation and Environments Agents

This type of agent utilizes trial-and-error in simulated environments for training, which is essential for tasks where physical trials are impractical or risky. Typically, research on these agents involves simulation platforms for navigation, object manipulation, and human-agent interaction. (Tsoi et al., 2022; Deitke et al., 2020; Kolve et al., 2017; Wang

et al., 2023d; Mees et al., 2022; Yang et al., 2023a; Ehsani et al., 2021; Savva et al., 2019; Szot et al., 2021; Puig et al., 2018; Carroll et al., 2019; Puig et al., 2023; Li et al., 2021; Srivastava et al., 2022; Mittal et al., 2023; Zhong et al., 2023; Liu & Negrut, 2021; Saito et al., 2023)

4.3. AR/VR/Mixed-reality Agents

These agents enhance the creation of interactive content in gaming and VR, enabling users to author their own experiences with advanced AI models and tools (Chen et al., 2021; Mao et al., 2022; Wake et al., 2023a). In some cases, agents are designed to assist in creating characters, environments, and objects in virtual worlds, streamlining the creation process and enabling dynamic generation and interaction within XR settings. (Huang et al., 2023b)

4.4. Knowledge and Logical Inference Agents

This agent focuses on applying knowledge and logical reasoning, integrating implicit and explicit knowledge sources for 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 & Davis, 2019; Gao et al., 2020; Wang et al., 2023a; Chen et al., 2020; Park et al., 2023a).

4.5. Agents for Emotional Reasoning

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)

4.6. Neuro-symbolic Agents

Neuro-symbolic Agents operate on a hybrid system of neurons and symbols, solving problems stated in natural language by capturing discrete symbolic structural information. (Chen et al., 2020; Park et al., 2023a)

These categories of Agents emphasize the importance of using multimodal information to take beneficial actions from their respective aspects. This indicates the necessity for Agents to possess high recognition capabilities for both language and images, thereby strongly suggesting the effectiveness of leveraging LLMs/VLMs.

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 representative sub-tasks that Agent AI is applied.

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 LLMs/VLMs have been applied, and share findings from our most recent studies.

Multimodal Systems. Recent research focuses on developing end-to-end systems incorporating LLM/VLM technologies as encoders for input information, guiding robotic actions based on linguistic instructions and visual cues (Jiang et al., 2022; Brohan et al., 2023; 2022; Li et al., 2023g; Ahn et al., 2022b; Shah et al., 2023b; Li et al., 2023d).

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., 2023d). For skill training, LLMs/VLMs 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). 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., 2023b). 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). Advanced technologies enable robots to navigate in challenging environments using object names (Chaplot et al., 2020; 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).

5.2. Agents for Gaming

Games provide a unique sandbox to test the agentic behavior of LLMs/VLMs, 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 LLMs, 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 LLMs/VLMs, can analyze and learn from human behavior, providing more human-like interactions and increasing realism and immersion.

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. In modern games, providing vast open-world environments necessitates the use of procedural or AI-driven techniques for automated terrain generation. Agent AI, utilizing LLMs/VLMs, 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 LLMs/VLMs 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 (Li et al., 2023b). 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 Multimodality

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 and 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 (Liu et al., 2023; Li et al., 2023e; 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., 2023f; 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.

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 to address common issues in large-scale models, such as hallucinations and biases in their outputs. the development of agents paradigm with integrated modalities (audio, image, text, sensor inputs), aiming to enhance their recognition and response capabilities for a wide variety of applications.

General-purpose end-to-end systems. the development of end-to-end models that are trained with large-scale data, seeking to create versatile and adaptable AI solutions.

Methodologies for grounding modalities. integrating information across various modalities, enhancing the coherence and efficacy of data processing.

Intuitive human interface. the development of effective and meaningful interaction between humans and agents.

Taming LLM/VLMs. exploring new approaches to 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 this, 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).

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.

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 participant's individual perspectives, 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. Ethical Discussion and Consideration

Multimodal 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 LLM and VLM agents, 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.

9. Conclusion

This 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 will bring together researchers in the field of agent AI with expertise in large foundation model based embodied modules in exploring the holistic intersections. By leveraging the collective expertise of agent paradigm, agent foundation model, agent infrastructure, and agent system from various AI disciplines. This paper aims to not only advance scientific interactive understanding of Agent AI, but also to discuss the embodied agent at the frontier of novel holistic intelligence research and helps us position ourselves to capitalize on emerging foundational models.

10. Impact Statement

One of the main goals of the 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.

Our 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 LLMs and VLMs 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 frame-

work 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.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.-H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M., and Zeng, A. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022a.

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

Allevato, A., Short, E. S., Pryor, M., and Thomaz, A. Tunenet: One-shot residual tuning for system identification and sim-to-real robot task transfer. In *Conference on Robot Learning*, pp. 445–455. PMLR, 2020.

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

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., and Wijmans, E. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cai, W., Huang, S., Cheng, G., Long, Y., Gao, P., Sun, C., and Dong, H. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. *arXiv preprint arXiv:2309.10309*, 2023.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Chaplot, D. S., Gandhi, D. P., Gupta, A., and Salakhutdinov, R. R. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.

Chen, G., Zheng, Y.-D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., and Wang, L. Videollm: Modeling video sequence with large language models, 2023.

Chen, K., Huang, Q., Palangi, H., Smolensky, P., Forbus, K. D., and Gao, J. Mapping natural-language problems to formal-language solutions using structured neural representations. In *ICML 2020*, July 2020.

Chen, K., Huang, Q., McDuff, D., Gao, X., Palangi, H., Wang, J., Forbus, K., and Gao, J. Nice: Neural image commenting with empathy. In *EMNLP 2021*, October 2021. URL https://www.microsoft.com/en-us/research/publication/nice-neural-image-commenting-with-empathy/.

Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3164–3174, 2020.

Dorbala, V. S., Sigurdsson, G., Piramuthu, R., Thomason, J., and Sukhatme, G. S. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*, 2022.

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

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Ehsani, K., Han, W., Herrasti, A., VanderBilt, E., Weihs, L., Kolve, E., Kembhavi, A., and Mottaghi, R. Manipulathor: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4497–4506, 2021.

Gadre, S. Y., Wortsman, M., Ilharco, G., Schmidt, L., and Song, S. 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*, pp. 23171–23181, 2023.

Gao, J., Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., and Shum, H.-Y. Robust conversational ai with grounded text generation. *arXiv preprint arXiv:2009.03457*, 2020.

Gao, J., Xiong, C., Bennett, P., and Craswell, N. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*, 2022.

Gervet, T., Chintala, S., Batra, D., Malik, J., and Chaplot, D. S. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991, 2023.

Granic, I., Lobel, A., and Engels, R. C. The benefits of playing video games. *American psychologist*, 69(1):66, 2014.

Guimarães, R. L., de Oliveira, A. S., Fabro, J. A., Becker, T., and Brenner, V. A. Ros navigation: Concepts and tutorial. *Robot Operating System (ROS) The Complete Reference (Volume 1)*, pp. 121–160, 2016.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Hensel, L. B., Yongsatianchot, N., Torshizi, P., Minucci, E., and Marsella, S. Large language models in textual analysis for gesture selection. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, pp. 378–387, 2023.

Ho, D., Rao, K., Xu, Z., Jang, E., Khansari, M., and Bai, Y. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10920–10926. IEEE, 2021.

Huang, C., Mees, O., Zeng, A., and Burgard, W. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615. IEEE, 2023a.

Huang, Q., Park, J. S., Gupta, A., Bennett, P., Gong, R., Som, S., Peng, B., Mohammed, O. K., Pal, C., Choi, Y., et al. Ark: Augmented reality with knowledge interactive emergent ability. *arXiv preprint arXiv:2305.00970*, 2023b.

Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9118–9147. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/huang22a.html.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., and Ichter, B. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022b.

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

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., and Fan, L. Vima: General robot manipulation with multimodal prompts. *arXiv*, 2022.

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

Katara, P., Xian, Z., and Fragkiadaki, K. Gen2sim: Scaling up robot learning in simulation with generative models. *arXiv preprint arXiv:2310.18308*, 2023.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*, 2016.

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

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

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.

Li, B., Wu, P., Abbeel, P., and Malik, J. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*, 2023a.

Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023b.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023c.

Li, J., Gao, Q., Johnston, M., Gao, X., He, X., Shakiah, S., Shi, H., Ghanadan, R., and Wang, W. Y. Mastering robot manipulation with multimodal prompts through pretraining and multi-task fine-tuning. *arXiv preprint arXiv:2310.09676*, 2023d.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023e.

Li, K., He, Y., Yi, W., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023f.

Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023g.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. In *arXiv preprint arXiv:2209.07753*, 2022.

Liang, X., Ma, L., Guo, S., Han, J., Xu, H., Ma, S., and Liang, X. Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation. *arXiv preprint arXiv:2306.10322*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context. *Proceedings of ECCV*, 2014.

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

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023.

Ma, Y. J., Liang, W., Wang, G., Huang, D.-A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., and Anandkumar, A. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2023.

Mao, R., Liu, Q., He, K., Li, W., and Cambria, E. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 2022.

Marcus, G. and Davis, E. *Rebooting AI: Building artificial intelligence we can trust*. Pantheon, 2019.

Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., and Garcia-Rodriguez, J. Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *Virtual Reality*, 24:271–288, 2020.

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

Meta Fundamental AI Research (FAIR) Diplomacy Team, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J. L., Singh, R., Guo, Y., Mazhar, H., et al. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023.

Müller, M., Casser, V., Lahoud, J., Smith, N., and Ghanem, B. Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision*, 126:902–919, 2018.

Ni, Z., Deng, X.-X., Tai, C., Zhu, X.-Y., Wu, X., Liu, Y.-J., and Zeng, L. Grid: Scene-graph-based instruction-driven robotic task planning. *arXiv preprint arXiv:2309.07726*, 2023.

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

Parakh, M., Fong, A., Simeonov, A., Gupta, A., Chen, T., and Agrawal, P. Human-assisted continual robot learning with foundation models. *arXiv preprint arXiv:2309.14321*, 2023.

Park, J. S., Hessel, J., Chandu, K., Liang, P. P., Lu, X., West, P., Huang, Q., Gao, J., Farhadi, A., and Choi, Y. Multimodal agent – localized symbolic knowledge distillation for visual commonsense models. In *NeurIPS 2023*, October 2023a.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023b.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., and Torralba, A. Virtualhome: Simulating household activities via programs. In *2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8494–8502, 2018.

Puig, X., Undersander, E., Szot, A., Cote, M. D., Yang, T.-Y., Partsey, R., Desai, R., Clegg, A. W., Hlavac, M., Min, S. Y., et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

Ramakrishnan, S. K., Chaplot, D. S., Al-Halah, Z., Malik, J., and Grauman, K. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18890–18900, 2022.

Raman, S. S., Cohen, V., Paulius, D., Idrees, I., Rosen, E., Mooney, R., and Tellex, S. Cape: Corrective actions from precondition errors using large language models. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

Rao, K., Harris, C., Irpan, A., Levine, S., Ibarz, J., and Khansari, M. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11157–11166, 2020.

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

Saito, D., Sasabuchi, K., Wake, N., Takamatsu, J., Koike, H., and Ikeuchi, K. Task-grasping from a demonstrated human strategy. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 880–887, 2022. doi: 10.1109/Humanoids53995.2022.10000167.

Saito, D., Sasabuchi, K., Wake, N., Kanehira, A., Takamatsu, J., Koike, H., and Ikeuchi, K. Constraint-aware policy for compliant manipulation, 2023.

Sasabuchi, K., Saito, D., Kanehira, A., Wake, N., Takamatsu, J., and Ikeuchi, K. Task-sequencing simulator: Integrated machine learning to execution simulation for robot manipulation. *arXiv preprint arXiv:2301.01382*, 2023.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

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

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

Shah, S., Dey, D., Lovett, C., and Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pp. 621–635. Springer, 2018.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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*, 2018.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Sontakke, S. A., Zhang, J., Arnold, S. M., Pertsch, K., Bıyık, E., Sadigh, D., Finn, C., and Itti, L. Roboclip: One demonstration is enough to learn robot policies. *arXiv preprint arXiv:2310.07899*, 2023.

Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K. E., Lian, Z., Gokmen, C., Buch, S., Liu, K., et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pp. 477–490. PMLR, 2022.

Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., and Batra, D. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Teshima, H., Wake, N., Thomas, D., Nakashima, Y., Kawasaki, H., and Ikeuchi, K. Deep gesture generation for social robots using type-specific libraries. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8286–8291. IEEE, 2022.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. 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)*, pp. 23–30. IEEE, 2017.

Tsoi, N., Xiang, A., Yu, P., Sohn, S. S., Schwartz, G., Ramesh, S., Hussein, M., Gupta, A. W., Kapadia, M., and Vázquez, M. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4):11047–11054, 2022.

Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., and Ikeuchi, K. Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.11753*, 2023a.

Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., and Ikeuchi, K. Gpt models meet robotic applications: Co-speech gesturing chat system. *arXiv preprint arXiv:2306.01741*, 2023b.

Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., and Ikeuchi, K. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023c.

Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., and Ikeuchi, K. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*, 11:95060–95078, 2023d. doi: 10.1109/ACCESS.2023.3310935.

Wang, B., Huang, Q., Deb, B., Halfaker, A. L., Shao, L., McDuff, D., Awadallah, A., Radev, D., and Gao, J. Logical transformers: Infusing logical structures into pre-trained language models. In *Proceedings of ACL 2023*, May 2023a.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023b.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., and Qiao, Y. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2023c.

Wang, Y., Xian, Z., Chen, F., Wang, T.-H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., and Gan, C. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023d.

Wang, Z., Cai, S., Liu, A., Ma, X., and Liang, Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023e.

Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Jiang, C., Tan, H., Kang, J., Zhang, Y., Zhou, K., et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023a.

Yang, K., Ji, S., Zhang, T., Xie, Q., and Ananiadou, S. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*, 2023b.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models, 2023.

Ye, Y., You, H., and Du, J. Improved trust in human-robot collaboration with chatgpt. *IEEE Access*, 2023.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 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*, 2014.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Yu, W., Gileadi, N., Fu, C., Kirmani, S., Lee, K.-H., Arenas, M. G., Chiang, H.-T. L., Erez, T., Hasenclever, L., Humplik, J., et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.

Zhang, S., Song, X., Bai, Y., Li, W., Chu, Y., and Jiang, S. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15130–15140, 2021.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., and Qin, B. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.

Zhong, Z., Cao, J., Gu, S., Xie, S., Gao, W., Luo, L., Yan, Z., Zhao, H., and Zhou, G. Assist: Interactive scene nodes for scalable and realistic indoor simulation. *arXiv preprint arXiv:2311.06211*, 2023.

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

Zhou, H., Ding, M., Peng, W., Tomizuka, M., Shao, L., and Gan, C. Generalizable long-horizon manipulations with large language models. *arXiv preprint arXiv:2310.02264*, 2023b.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017a.

Zhu, S., Kimmel, A., Bekris, K. E., and Boularias, A. Fast model identification via physics engines for data-efficient policy search. *arXiv preprint arXiv:1710.08893*, 2017b.