Position: Foundation Agents as the Paradigm Shift for Decision Making

Xiaoqian Liu [1 2] **Xingzhou Lou** [3 2] **Jianbin Jiao** [4] **Junge Zhang** [2]

Abstract

Decision making demands intricate interplay between perception, memory, and reasoning to discern optimal policies. Conventional approaches to decision making face challenges related to low sample efficiency and poor generalization. In contrast, foundation models in language and vision have showcased rapid adaptation to diverse new tasks. Therefore, we advocate for the construction of foundation agents as a transformative shift in the learning paradigm of agents. This proposal is underpinned by the formulation of foundation agents with their fundamental characteristics and challenges motivated by the success of large language models (LLMs). Moreover, we specify the roadmap of foundation agents from large interactive data collection or generation, to self-supervised pretraining and adaptation, and knowledge and value alignment with LLMs. Lastly, we pinpoint critical research questions derived from the formulation and delineate trends for foundation agents supported by real-world use cases, addressing both technical and theoretical aspects to propel the field towards a more comprehensive and impactful future.

1. Introduction

Human life is all about making decisions. Intelligent agents have been developed to help humans with real-world decision making, such as traffic control (Liang et al., 2019), energy management (Nakabi & Toivanen, 2021), and drug discovery (Zhou et al., 2019). Prevalent paradigms to train agents include reinforcement learning (RL), imitation learning (IL), planning and search as well as optimal control. Recent advances in these algorithms have achieved superhuman performance in mastering the game of Go (Silver et al., 2017), playing video games (Schwarzer et al., 2023), and robotic locomotion and manipulation (Brohan et al., 2022). However, traditional approaches to decision making exhibit limited sample efficiency and poor generalization. For instance, expert systems heavily rely on human knowledge and manual crafting, and conventional RL methods necessitate agent training from scratch for each task.

In contrast, foundation models (Bommasani et al., 2021) in language and vision achieve rapid adaptation to a wide variety of tasks with minimal fine-tuning or prompting. These models, pretrained on vast and diverse datasets, demonstrate unprecedented capabilities in understanding and generating text (Brown et al., 2020; OpenAI, 2023), image (Dosovitskiy et al., 2020; Bai et al., 2023) and some multimodalities (Reed et al., 2022; Lu et al., 2023). Therefore, this position paper argues that **foundation agents as generally capable agents across physical and virtual worlds, will be the paradigm shift for decision making**, akin to LLMs as general-purpose language models to solve linguistic and knowledge-based tasks.

To arrive at this position, we first identify three fundamental characteristics of foundation agents: (1) *a unified representation* of variables involved in decision process, including state-action spaces, feedback signals (e.g., rewards or goals) and environment dynamics, (2) *a unified policy interface* across tasks and domains from robotics and game play to healthcare and beyond, and (3) *interactive decision-making* in physical and virtual worlds by reasoning about behaviors, handling environment stochasticity and uncertainty, and potentially navigating competitive or cooperative multi-agent scenarios. These characteristics constitute the uniqueness and challenges for foundation agents[1], empowering them with multi-modality perception, multi-task and cross-domain adaptation as well as few- or zero-shot generalization. Particularly, foundation agents exhibit enhanced credit assignment and planning in scenarios requiring long-horizon reasoning (Wei et al., 2022; Yao et al., 2022; 2023) and involving sparse rewards or partial observability (Meng

[1]School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing, China [2]Institute of Automation, Chinese Academy of Sciences, Beijing, China [3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China [4]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Junge Zhang <jgzhang@nlpr.ia.ac.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Distinctions from task-specific agents and recent LLM-based agents can be found in Appendix A.

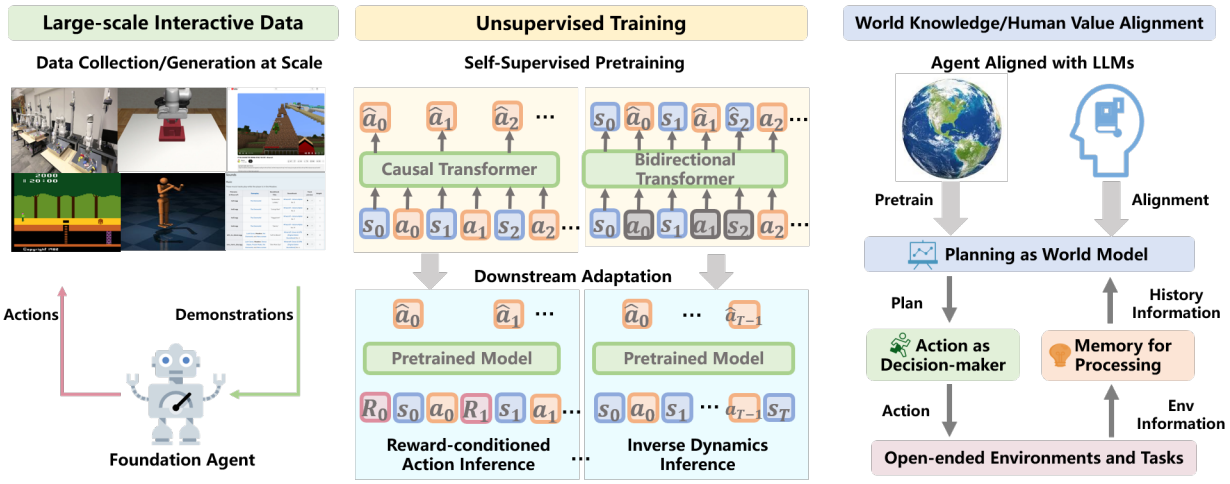

Figure 1. Roadmap of foundation agents. **Left:** Large interactive data collection or generation scales up foundation agents in open-ended physical and virtual worlds. **Middle:** Self-supervised pretraining leverages the flexible and robust architecture of Transformer (Vaswani et al., 2017) based on autoregressive or masked (in gray) modeling. During adaptation, the pretrained model play various roles in decision making, such as policy initialization (Meng et al., 2021; Yang & Nachum, 2021) and dynamics model (Wu et al., 2023; Brandfonbrener et al., 2023). Distinct colors denote different variables within trajectories. **Right:** LLMs as a part of foundation agents composed of planning, memory and action modules and act as world models, information processors, and decision-makers, respectively.

et al., 2021; Guo et al., 2023). Such unique set of characteristics and capabilities enable foundation agents to improve sample efficiency and generalization, showing their versatility in navigating diverse contexts of decision making.

With the formulation of foundation agents, we specify their roadmap in Figure 1. Firstly, *large-scale interactive data* can be collected from the internet (e.g., YouTube videos, tutorials, audios, etc) and physical environments, or generated through real-world simulators (Yang et al., 2023a; OpenAI, 2024; Bruce et al., 2024) for supervision and scale-up. Secondly, *pretrain in an unsupervised manner* from large, unlabeled, and probably suboptimal interactive data for decision-related knowledge representation learning and downstream adaptation with knowledge reasoning. Thirdly, align with LLMs to integrate *world knowledge and human values* into foundation agents for reasoning, generalization, and interpretability. This roadmap is motivated by three key ingredients we observe for the success of LLMs: (1) leveraging internet-scale text data to absorb broad knowledge described by languages, (2) self-supervised pretraining to learn unified text representation (Devlin et al., 2018) and task interface (Raffel et al., 2020) through generative modeling, and (3) safety and preference alignment (Ouyang et al., 2022; Rafailov et al., 2024) to fulfill user commands.

For foundation agents, however, it is challenging to reproduce the roadmap of LLMs. Firstly, the broad information in physical and virtual worlds are low-level details instead of high-level abstractions expressed by languages, posing challenges to a unified representation. Therefore, we discuss the morphology of foundation agents (e.g, a unified or compositional foundation model) for structuring diverse environments and tasks in (§6.1). Secondly, the large domain gap between different decision making scenarios makes it difficult to develop a unified policy interface, while linguistic tasks can be simply unified by text generation. We thus posit that theoretical guarantees for policy optimization with foundation agents may address this issue in (§6.2). Thirdly, while language and vision models focus on understanding and generating content, foundation agents are involved in dynamic process of choosing optimal actions based on complex environment information. Open-ended tasks and environments are then another critical research question for foundation agents as discussed in (§6.3). Finally, we showcase some examples supported by recent work and a case study towards building a potential foundation agent in real-world decision making.

2. Preliminaries

2.1. Decision Making Formalism

Decision making refers to the process of making a series of decisions to achieve a goal in continuous time, based on previous actions and observations while considering possible future states and rewards. The process can be simplified as a markov decision process (MDP) $M =< S, A, T, R, \gamma >$, where S is the state space, A the action space, T the dynamics transition function $T : S \times A \times S \to [0, 1)$, R the reward function $R : S \times A \times S \to \mathbb{R}$, and $\gamma \in [0, 1)$ is a discount factor for calculating discounted cumulative rewards. When the underlying state is not accessible (e.g,

a video game), the process can be modified as a partially observable MDP $M =< S, A, T, R, \mathcal{O}, E >$, where $\mathcal{O}$ is the observation space, and $E(o|s)$ denotes the observation emission function. Solutions to MDP typically involve RL to learn an optimal policy that maximizes the expected discounted cumulative rewards $\mathbb{E}[\sum_{t=0}^{T} \gamma^t r_{t+1}]$ through trial and error within an environment, or IL to learn an expert's policy from expert demonstrations in a supervised learning manner. However, real-world decision making is more complex and not limited to MDP, such as stock market dynamics, epidemiological modeling, and sequential games with incomplete information (e.g, Poker and Bridge).

2.2. Self-Supervised Learning for RL

To improve sample efficiency and generalization, conventional self-supervised learning for RL involves representation learning of separate RL components, such as action representation (Gu et al., 2022), reward representation (Ma et al., 2022), policy representation (Tang et al., 2022), and environment or task representation (Sang et al., 2022). Recently, by formulating RL as a sequence modeling problem (Chen et al., 2021; Janner et al., 2021), the representation of various RL components can be simultaneously learned via trajectory optimization (Lee et al., 2022; Liu et al., 2022; Carroll et al., 2022; Sun et al., 2023; Wu et al., 2023; Boige et al., 2023). Such a formulation draws upon the simplicity and scalability of Transformer (Vaswani et al., 2017), and benefits representation learning of knowledge concerning different aspects of decision process.

2.3. Large Language Models and Agents

Large language models (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022) are pretrained on internet text and fine-tuned on human instructions and preferences. Through this process, LLMs acquire extensive world knowledge and are aligned with human values (OpenAI, 2023; Anthropic, 2023; Rafailov et al., 2024; Touvron et al., 2023). State-of-the-art LLMs demonstrate convincing capabilities in not only natural language processing (Völske et al., 2017; Singhal et al., 2022; Jiao et al., 2023), but also tasks requiring strong reasoning abilities like coding (Nijkamp et al., 2022) and text-based games (Shridhar et al., 2020; Liu et al., 2023b). As a source of world knowledge and human values, we argue that LLMs enable foundation agents to align with the world model and human society created by languages, enhancing their reasoning and planning capabilities.

3. Learning from Large-Scale Interactive Data

Large-scale interactive data is an essential component for building foundation agents, akin to the significance of internet text and image data for foundation models in language and vision. In this section, we first demonstrate how a potential foundation agent can be trained via offline IL when large, multi-modal and multi-task demonstrations are available. We then discuss the potential use of data generation systems or real-world simulators for training foundation agents at scale. Finally, we identify the constraints and alternatives to offline RL or IL methods in establishing foundation agents, as well as the limitations of real-world simulators realized by video generation models.

3.1. Offline IL from Large Demonstrations

Inspired by sequence modeling of RL problems, recent work attempt to train a generalist agent on extensive interactive demonstrations simply via offline IL. For example, by imitating multi-modal demonstrations, Behavior Transformer (Shafiullah et al., 2022) leverages a modified Transformer with action discretization and a multi-task action correction to capture the modes present in large behavior data. Gato (Reed et al., 2022) unifies multi-modal and multi-task expert episodes into sequences through autoregressive sequence modeling, and is capable of diverse tasks and domains from robotic manipulation and Atari play to visual question answering. Similarly, Unified-IO 2 (Lu et al., 2023) scales autoregressive multi-modal model across vision, language, audio, and action data, unifying different modalities into a shared semantic space. With a single encoder-decoder Transformer, it learns varieties of skills through fine-tuning with prompts and augmentations. However, collecting demonstrations via humans or online RL algorithms is costly and time-consuming, which is not applicable in real-world decision making.

3.2. Potentials of Real-World Simulators

Generative models as data generation systems (Mandlekar et al., 2023) or real-world simulators (Yang et al., 2023a; OpenAI, 2024; Bruce et al., 2024) open another path for training foundation agents through learning behaviors from generated data. For instance, RoboGen (Wang et al., 2023c) realizes a propose-generate-learn cycle to scale up robot learning via generative simulation, of which policy learning algorithms are not restricted to IL but also involve motion planning or trajectory optimization. Moreover, video generation models such as Sora (OpenAI, 2024) and Genie (Bruce et al., 2024) can be a general-purpose simulator of interactive decision making scenarios in physical and virtual worlds. These models, by unifying the representation of visual worlds (Yang et al., 2024), generate diverse and open-ended training environments across tasks and domains for scaling up the foundation agent in a never-ending curriculum of new and generated worlds. The foundation agent pretrained with these generated environments learns to reason about behaviors, handle environment stochasticity and uncertainty, and navigate competitive or cooperative multi-agent settings, thus enhancing adaptability and generaliza-

tion in unseen scenarios. This could be particularly valuable for domains where real-world interactive data is scarce or expensive to obtain, such as robotics and self-driving.

3.3. Discussion

Despite the preliminary efforts, real-world interactive data scales far beyond internet text or visual data, making it impossible to solely rely on offline RL or IL to train foundation agents. Alternative approaches, such as self-supervised (unsupervised) pretraining, can be utilized to harness large and unlabeled interactive data. Moreover, current generalist agents remain small in model size compared to LLMs, but already consume significant computational resources even when trained on a single domain. We provide a summary of computational requirements of recently proposed generalist agents and a case we build in Appendix B.

In addition, although we posit that universal video generation models (e.g., Sora and Genie) hold promise for enhancing the training of foundation agents, they are not the whole story. Firstly, not all real-world decision making tasks can be adequately represented by video, such as situations when visual and textual information is not available like wireless communication and grid management. Secondly, the video generation task may not be a unified policy interface, since simply predicting next frame may not induce a reasonable agent behavior. Specifically, video generation objectives such as generating realistic video sequences given a fixed context, cannot be directly compatible with decision making objectives that optimize policies given feedback signals. Thirdly, as the real world is much more complicated than simulators, the training of foundation agents should be grounded in simplified world such as a state formed as a result of an action. Then the agent can be trained on high-level results rather than low-level pixels and videos, which remains an open question for future work.

4. Self-Supervised Pretraining and Adaptation

Interactive data captures various aspects of information in the decision process, including state transitions, state-action causality and state or action values if rewards are provided. These decision-related knowledge should be learned during pretraining and transferred to downstream inference tasks so as to improve sample efficiency and generalization of agents. Similar to foundation models in language and vision, we posit that the pretraining and adaptation pipeline can be considered as knowledge representation learning and reasoning of foundation agents. In this section, we highlight strategies in self-supervised pretraining for decision making, and discuss the potential of pretraining and adaptation in building foundation agents.

4.1. Self-Supervised Pretraining

Self-supervised (unsupervised) pretraining for decision making allows foundation agents to learn without reward signals and encourages the agent to learn from suboptimal offline datasets. This is particularly applicable when large, unlabeled data can be easily collected from internet or real-world simulators. Specifically, given a sequence of trajectories $\tau = \{(s_1, a_1, s_2, a_2, , ..., s_T, a_T)\}^N$, self-supervised pretraining aims to learn a representation function $g : \mathbb{T} \in \mathbb{R}^d \rightarrow \mathbb{Z} \in \mathbb{R}^m$ $(m \ll d)$ to distill valuable knowledge from trajectory data for downstream inference. The knowledge can be temporal information about the same modality (e.g., $s_t \rightarrow s_{t+i}$), causal information between different modalities (e.g., $\pi(a|s)$), as well as dynamics $P(s'|s, a)$ and reward $R(s, a)$ information.

Generally, there are two steps in pretraining foundation agents based on Transformer architecture. The first step is to learn embeddings of trajectory data. Specifically, the tokenization of trajectory sequences comprises three components:(1) *trajectory encoding* that transforms raw trajectory inputs into a common representation space, (2) *timestep encoding* that captures absolute or relative positional information, and (3) *modality encoding* to disambiguate between different modalities in trajectories (Wu et al., 2023). Particularly, two levels of tokenization granularity have been studied: (1) discretization at the level of modalities (Sun et al., 2023; Wu et al., 2023), and (2) discretization at the level of dimensions (Reed et al., 2022; Boige et al., 2023).

The second step is devising self-supervised pretraining objectives to discover the underlying structure and semantics of trajectory data for knowledge representation learning. Table 1 shows pretraining objectives for decision making in domains including control (Sun et al., 2023), navigation (Carroll et al., 2022), and game play (Lee et al., 2022). These objectives are mainly inspired by autoregressive (Radford et al., 2019; Brown et al., 2020) or masked (Devlin et al., 2018; He et al., 2022) prediction in language and vision model pretraining. The primary pretraining objective for decision making involves learning control information by predicting the next action in an autoregressive way. This objective is further modified by conditioning on different variables within trajectories, such as reward or value signals (Chen et al., 2021; Lee et al., 2022), next state (observation) information (Sun et al., 2023), or latent future sub-trajectory information (Xie et al., 2023). As an alternative, random masking (Liu et al., 2022) learns the context of trajectory data by filling in missing information. Various masking schemes conditioned on different RL components are designed. For example, reward-conditioned random mask prediction (Carroll et al., 2022) recovers masked trajectory segments conditioned on the first-step return, while (Sun et al., 2023) only recovers masked actions conditioned on

Table 1. Self-supervised pretraining objectives for decision making with Transformer. P_θ is the prediction of the pretrained model. $\tau_{0:t-1}$ denotes previous trajectories before timestep t. s_t is the state or observation at timestep t and a_t the action at timestep t. R_t is the return-to-go (target return for the rest of the input sequence) at timestep t, while $\hat{R}_t$ is the predicted future value and r_t the future reward. G_{t+i} refers to a sub-goal or the ultimate goal at future timesteps, and z denotes the encoded future trajectory in the same sequence length as the input trajectory. Note that the conditional variables in each loss function are encoded by either a causal Transformer (Brown et al., 2020) for autoregressive modeling (top rows) or a bidirectional Transformer (Devlin et al., 2018) for masked modeling (bottom rows).

Pretraining Objective	Loss Function
Next action prediction	$-logP_\theta(a_t \| \tau_{0:t-1}, s_t)$
Reward-conditioned action prediction	$-logP_\theta(a_t \| \tau_{0:t-1}, s_t, R_t)$
Future value and reward prediction	$-logP_\theta(a_t, \hat{R}_t, r_t \| \tau_{0:t-1}, s_t)$
Goal-conditioned action prediction	$-logP_\theta(a_t \| \tau_{0:t-1}, s_t, G_{t+i})$
Future-conditioned action prediction	$-logP_\theta(a_t \| \tau_{0:t-1}, s_t, z)$
Forward dynamics prediction	$-logP_\theta(s_t \| \tau_{0:t-1})$
Inverse dynamics prediction	$-logP_\theta(a_t \| s_t, s_{t+1})$
Random masking prediction	$-logP_\theta(masked(\tau) \| unmasked(\tau))$
Reward-conditioned random mask prediction	$-logP_\theta(masked(\tau) \| unmasked(\tau), R_0)$
Random masked hindsight prediction	$-logP_\theta(masked(a) \| unmasked(\tau), a_T)$
Random autoregressive mask prediction	$-logP_\theta(masked(\tau, a_T) \| unmasked(\tau))$

the final-step action to capture global temporal relations for multi-step control. Combining masked and autoregressive modeling, (Wu et al., 2023) constrains the last variable in trajectories to be masked to force the pretrained model to be causal at inference time. Moreover, contrastive prediction objective (Stooke et al., 2021; Schwarzer et al., 2021; Yang & Nachum, 2021; Cai et al., 2023) has been commonly used in self-supervised pretraining to learn state representation via a contrastive loss, which can benefit dynamics learning.

4.2. Downstream Adaptation

During adaptation, extensive decision-related knowledge acquired from pretraining is transferred to downstream tasks via fine-tuning or prompting. The knowledge transfer facilitates the optimization of learning objectives $f_\theta(z)$ in downstream inference, such as the value function $V(\pi)$ or $Q(s, a)$, policy, dynamics and reward functions. The optimized objectives empowered by knowledge reasoning can finally improve sample efficiency and generalization compared to learning from scratch in traditional RL.

Generally, there are two cases requiring fine-tuning: (1) when the pretraining data is a *mix* of a small proportion of near-expert data and a large proportion of exploratory trajectories, and (2) when the pretraining objective *significantly differs* from the inference objective of downstream decision making tasks. For instance, traditional RL aims to maximize cumulative rewards according to a specified reward function, whereas self-supervised pretraining tasks are usually reward-free. In such cases, RL algorithms are demanded for fine-tuning policies in online or offline settings (Yang & Nachum, 2021; Lee et al., 2022; Sun et al., 2023; Cai et al., 2023; Boige et al., 2023).

In addition to fine-tuning, prompting directly adapts the pretrained model to downstream inference without altering or introducing any parameters. It concerns: (1) *aligning* pretraining objectives with downstream inference objectives, and (2) *prompting* the model with interactive demonstrations (Xu et al., 2022; Reed et al., 2022; Laskin et al., 2022; Wei et al., 2023). For example, using a random masking pretraining objective with a variable mask ratio for different goals, (Liu et al., 2022) achieves zero-shot generalization to goal-reaching tasks. This is attributed to the natural alignment of the masked pretraining objective with goal-reaching scenarios, where the model is required to recover masked actions based on remaining states.

4.3. Discussion

Previous attempts at self-supervised RL pretraining have been mostly limited to a single task (Yang & Nachum, 2021), or performing pretraining and fine-tuning within the same task (Schwarzer et al., 2021). This is not generic and flexible for adapting to various decision making tasks. Therefore, we argue for multi-modal and multi-task self-supervised pretraining in the evolution of foundation agents. Particularly, self-supervised pretraining empowers foundation agents to acquire a nuanced understanding of large interactive data, laying a robust foundation for knowledge learning and reasoning in adaptation. However, challenges arise in optimizing foundation agents, such as determining the optimal granularity and representation of trajectory data. Furthermore, striking a balance between versatility and task specificity remains an ongoing challenge. Despite these problems, combining self-supervised pretraining and versatile adaptation strategies, underscores the potential of foundation agents in capable of robust performance across a broad array of

decision making scenarios.

5. Knowledge and Value Alignment via LLMs

The broad real-world knowledge and human values embedded in LLMs pave the way for foundation agents with improved reasoning, generalization and interpretability. In this section, we discuss the roles of LLMs in enhancing decision making realized by memory, planning and action modules. We then outline some major challenges and potential solutions to align foundation agents with LLMs.

5.1. Memory and Information Processing

Foundation agents aligned with LLMs are often equipped with a memory module. The memory module functions as a repository, storing both task-specific information and past interaction history that enables agents to retrieve and refine relevant historical data for future planning. Similar as human memory (Izquierdo et al., 1999), the agent's memory can be categorized into short-term and long-term components. Short-term memory encapsulates recent information within the context window, while long-term memory encompasses historical information stored in external storage.

After perceiving and processing environmental state, agents can decide what information to remember and what is crucial for current decision making that demands retrieval. For example, mastered skills (Wang et al., 2023a) are explicitly stored to prevent forgetting and retrieved as needed to enhance decision making efficiency. Further, through alignment with LLMs, foundation agents can refine their existing memories, generate new insights based on current knowledge, and assist future decision making (Park et al., 2023; Shinn et al., 2023).

5.2. Planning with World Models

The abundant world knowledge and human values embedded in LLMs enables LLMs to serve as world models (Hao et al., 2023), simulating transition dynamics and reward functions (Ma et al., 2023) for foundation agents to perform planning. The agent then is able to reason about current observations and historical information, decomposing a complicated task into a sequence of strategic plans or sub-goals. Few-shot prompting (Zhao et al., 2021) can be one efficient means to utilize LLMs for enhancing agent's understanding and planning by providing high-quality demonstrations in short-term memory. For example, Chain-of-Thought (CoT) (Wei et al., 2022) decomposes tasks into logically coherent sub-goals, and Tree-of-Thought (ToT) (Yao et al., 2023) enables agents to generate potential solutions, allowing them to select the most reliable one. Furthermore, ReAct (Yao et al., 2022) generates logically correct reasoning paths, improving the quality of generated plans by assessing correctness and logical coherence of previously generated actions.

Moreover, feedback from the environment (Yao et al., 2022; Huang et al., 2022; Wang et al., 2023a; Zhu et al., 2023; Wang et al., 2023d), LLM critics (Shinn et al., 2023; Madaan et al., 2023; Chen & Chang, 2023), or humans (Huang et al., 2022) can also be leveraged to improve the planning of foundation agents. These feedback enables agents to adjust behaviors and plans, addressing obstacles encountered during execution and thus improving task fulfillment.

5.3. Action and Decision Making

For ultimate action execution and interaction with real or simulated environments, an action module is required for foundation agents to interpret plans from the planning module, translate subgoals into executable actions, and combine atomic actions to form more complex structured actions.

Specifically, the action module specifies action goals and available actions. Action goals articulate intended objectives, such as movement (Zhu et al., 2023), communication (Zhang et al., 2023a), and code generation (Wang et al., 2023a). Clear goal descriptions enable agents to comprehend and generate reasonable actions based on observations. Available actions, on the other hand, define the agent's action space, which can be expanded by adding recently acquired skills into memory (Wang et al., 2023a; Zhu et al., 2023). Additionally, equipped with LLMs, foundation agents can possess the ability to call external APIs or tools for problem solving. HuggingGPT (Shen et al., 2023) exemplifies this by using LLMs as the controller to manage existing AI models from the internet via APIs to execute each sub-task. Further, by connecting to external databases (Hu et al., 2023; Zhu et al., 2023), the agent can acquire extra task-related knowledge via external tools and plugins created by domain experts (Ge et al., 2023) for reliable and creative decision making.

5.4. Discussion

The integration of LLMs into agent learning offers advantages over conventional RL and IL methods. Firstly, LLMs require significantly fewer training samples (Guo et al., 2023), as they leverage pretraining on internet-scale text, acquiring extensive real-world knowledge and demonstrating strong generalization abilities. Secondly, state-of-the-art LLMs are inherently human-aligned, enabling them to make decisions consistent with human preferences and values (Zhang et al., 2023a). Thirdly, by providing detailed reasoning for actions during planning, the interpretability of foundation agents aligned with LLMs addresses the challenge of deploying black-box AI models (Došilović et al., 2018) in real-world decision making.

However, a substantial challenge for aligning foundation

agents with LLMs lies in the hallucination of LLMs . Hallucination refers to the phenomenon where the language model generates false content that is misaligned with real-world knowledge (Zhang et al., 2023b). Since hallucination detection still remains an open problem, aligning foundation agents with LLMs amplifies the risk of inexplicable abnormal behaviors. Given the broad range of tasks assigned to foundation agents, particularly those that are safety-critical, the hallucination in LLMs could result in serious consequences during decision making. Therefore, it is crucial to remain vigilant about the potential risks of foundation agents aligned with LLMs, and propose solutions to this issue for a safe and responsible deployment in real-world decision making.

6. Trends for Foundation Agents

Derived from the formulation and challenges for foundation agents, some critical research questions remain to be solved. In this section, we propose these issues and potential solutions, and discuss the trends for foundation agents with examples in recent work as well as a case study in real-world decision making scenarios.

6.1. A Unified or Compositional Foundation Agent

One central challenge in building foundation agents stems from the substantial diversity among decision making tasks, involving variations in state and action spaces, environment dynamics, and reward functions. In contrast, the unification of tasks in language is facilitated by the text-to-text generation paradigm (Brown et al., 2020; OpenAI, 2023). To structure diverse environments and tasks, some attempts have been made along two primary directions: (1) the pursuit of a *unified* foundation model (Reed et al., 2022; Brohan et al., 2022), and (2) the exploration of *compositions* involving multiple existing foundation models (Zeng et al., 2022; Ajay et al., 2023) for decision making.

To establish a unified foundation model for decision making, a core requirement is the unified representation of RL components. This entails encoding states, actions, and rewards from diverse environments and tasks into standardized tokens through sequence modeling (Reed et al., 2022; Janner et al., 2021; Chen et al., 2021). The subsequent step involves transforming these tokens into a consistent data modality, such as text descriptions (Zhu et al., 2023), generated videos (Du et al., 2023), code (Wang et al., 2023a), or text-image pairs (Brohan et al., 2023). However, due to the multi-modal and heterogeneous characteristics of interactive data, uncertainties persist regarding the efficacy of tokenization methods in compressing raw trajectories into compact tokens. In addition, whether a single data modality can comprehensively represent interactive data and effectively convey its underlying information and knowledge warrants further investigation.

While a unified foundation model might offer a comprehensive solution, it could also lead to increased complexity and challenges in interpretability. Instead, we consider whether integrating existing foundations models (e.g., large language or vision models) with moderate decision models is sufficient to address most (if not all) of decision making tasks. The integration of existing foundation models allows for leveraging their domain-specific strengths but may introduce issues in harmonizing different modalities and functionalities. Additionally, a compositional foundation agent could inherit both merits and demerits from other foundation models. Therefore, striking a balance between these approaches requires careful evaluation of specific requirements and constraints of target tasks, offering an intriguing avenue for future research in foundation agents.

6.2. Policy Optimization with Foundation Agents

In traditional RL, theoretical guarantees underpin algorithmic advancements, such as the Bellman optimality update in temporal difference learning and dynamic programming. Recent theoretical analyses have shed light on the delusions in sequence models for interaction and control (Ortega et al., 2021), yet the optimization of a generalist policy with foundation agents still lacks a solid theoretical foundation.

Establishing theoretical foundations for policy optimization with foundation agents is a complex but crucial endeavor. Firstly, policies generated by foundation agents will directly affect changes in real-world applications, requiring rigorous theoretical guarantees to ensure their effectiveness, optimality, safety and robustness. Meanwhile, the theoretical foundations for policy optimization can in turn help to understand the principles underlying the agent's behavior and performance, thereby enabling the development of algorithms and methodologies for training foundation agents. Specifically, we posit that the theoretical foundations require an interdisciplinary view (e.g., control theory, RL, and optimization), and we outline some ideas here.

- **Define Pretraining and Task-Specific Objectives**: Unlike traditional RL objectives that focus on maximizing cumulative rewards, the optimization objectives for foundation agents involve both pretraining and task-specific objectives. Pretraining objectives address the unified representation of variables involved in decision process and the unified policy interface across tasks and domains. Task-specific objectives may be employed in fine-tuning phase, which have been established in many control, planning, and RL algorithms (Yang et al., 2023a; Wang et al., 2023c).
- **Understand the Interplay between Pretraining and Task-Specific Optimization**: Establish a mathematical framework to investigate how the pretraining phase shapes

the optimization landscape for task-specific policy optimization, and how different adaptation algorithms interact with the previously acquired knowledge. This could involve probing convergence properties, sample efficiency, and generalization bounds of foundation agents.

- **Extend Existing Theoretical Frameworks**: Adapt and extend existing theoretical frameworks from RL, control theory and optimization to accommodate the peculiarities of foundation agents. This may involve relaxing assumptions made in traditional RL and developing new notions of optimality for generalist policies. For example, (1) *extend the MDP framework* with generalized state spaces to incorporate multi-modal inputs and extended action spaces spanning different domains (e.g., physical movements and linguistic responses), (2) leverage the probabilistic representations and principled reasoning offered by *the control as inference* framework (Todorov, 2008) to derive optimal control policies and analyze their properties, and (3) ground policy optimization for foundation agents in *information-theoretic* principles, such as mutual information or entropy, to capture the agent's ability to influence the environment and gather informative experiences.

6.3. Open-Ended Tasks for Foundation Agents

In the evolution of foundation agents, a notable trend is the shift towards learning from open-ended tasks at scale (Team et al., 2021; Fan et al., 2022). Traditionally, RL agents are confined to a single and individual task (Silver et al., 2017; Schwarzer et al., 2023), limiting their applicability to massive dynamic scenarios. Contemporary approaches instead prioritize the scalability of generative models, especially via LLMs to effectively handle the complexities inherent in open-ended tasks (Wang et al., 2023a; Zhu et al., 2023). The open-ended agents, endowed with continuously learning capabilities, interact with environments or external tools (Shen et al., 2023; Hu et al., 2023; Ge et al., 2023), iteratively refining their decision process. This paradigm aligns seamlessly with RL principles, where agents acquire optimal strategies through iterative learning.

Despite the promising advances in the development of open-ended agents, several challenges persist particularly for open-ended tasks. Importantly, open-ended tasks do not have a predefined goal or endpoint, push the boundaries of agent adaptability, generalization, and continual learning, and require flexible adaptation, creativity, and the ability to discover new goals or objectives. Inspired by these features, we discuss the issues associated with foundation agent learning in open-ended tasks, along with insights into how these issues can be addressed.

- **Continual Learning and Adaptation**: The dynamic nature of open-ended tasks poses challenges in *model adaptability without catastrophic forgetting* of acquired skills and knowledge. Agents need to handle an unlimited variety of open-ended goals with training objectives and task distributions dynamically changing. Two lines of recent work attempt to address the issue by (1) leveraging pretrained language or vision model to convert world knowledge into actionable insights (Ahn et al., 2022; Driess et al., 2023), and (2) pretraining (or meta-learning) large transformer models from scratch on multi-tasks (Team et al., 2021; Fan et al., 2022; Bauer et al., 2023). However, these studies require mostly human-defined goals as input. Instead, (Wang et al., 2023b) formulates goal as energy function and learns without goal-conditioning via diffusion models. Nevertheless, it is also limited by human-designed functions for goals which could be replaced by neural networks in future work.
- **Unboundedness and Novelty Detection**: In open-ended tasks, the environment might be *infinite or continually evolving*, making it difficult to exhaustively explore or anticipate all possible states and outcomes. Agents need to deal with *novelty and unexpected situations* throughout their lifetime. Current research attempts to address this issue through novelty difficulty measures (Pinto et al., 2023), hierarchical RL (Vezhnevets et al., 2017) and self-initiated open world learning framework (Liu et al., 2023a), yet it still remains an open question.
- **Curriculum and Autotelic Learning**: Agents learning in open-ended tasks have to learn and explore *without explicit external rewards*. Existing studies apply procedural environment generation to create diverse and increasingly complex environments (Bauer et al., 2023), and intrinsic motivation (Forestier et al., 2022) and curiosity-driven learning to encourage exploration. Future work can consider automatically generate curriculum via LLMs (Wang et al., 2023a), universal video generation models (OpenAI, 2024; Bruce et al., 2024) or multi-modal models.
- **Creativity and Innovative Solutions**: Generating *novel and valuable solutions* is a hallmark of open-ended learning. This necessitates agents that can go beyond mere imitation or optimization, possibly by combining known concepts in unique ways or inventing entirely new ones. Potential solutions include (1) Quality-Diversity algorithms (Grillotti & Cully, 2022) to produce a diverse set of high-performing solutions, (2) neuroevolution methods (Lehman & Stanley, 2011) to produce innovative solutions, and (3) deep generative models to generate novel and sometimes unexpectedly creative outputs.

6.4. Case Studies in Real-World Decision Making

Autonomous Control. The progress in robotics continually enhances productivity and resource efficiency. Recently, robots with general capabilities have been developed towards foundation agents. For example. RT-1 (Brohan et al.,

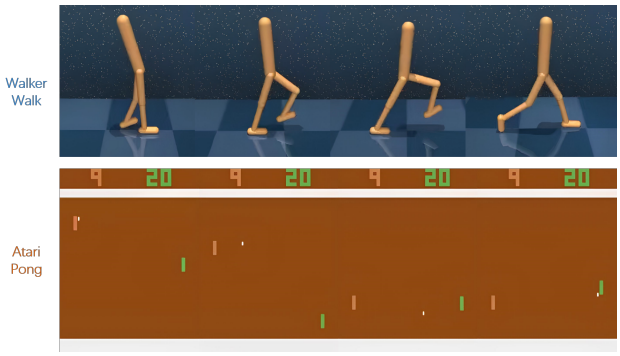

Figure 2. Visualized fine-tuning performance of our pretrained agent on unseen task and game. Frames are selected from the video recorded in the last evaluation episode. Total number of evaluation episodes is 50. Results are from one training seed.

2022) is a generalist and language-conditioned robotic agent, exhibiting zero-shot generalization to new tasks, environments and long-horizon scenarios by scaling on a large set of manipulation demonstrations (Padalkar et al., 2023). RT-2 (Brohan et al., 2023) further improves generalization and reasoning in response to user commands by co-fine-tuning pretrained vision-language models on demonstrations and internet vision-language tasks. Moreover, equipped with real-time human feedback, robots can also improve their high-level decision making iteratively in long-horizon tasks (Shi et al., 2024). We also conduct a case study in robotic locomotion and game play by jointly pretraining on 5 tasks in the DeepMind Control (DMC) suite (Tassa et al., 2018) and 5 games in Atari video play (Bellemare et al., 2013) through autoregressive modeling. Visualization of the fine-tuning performance on a new locomotion task and a new game is shown in Figure 2. Details of our case study can be found in Appendix C.

In addition to robotics and game play, self-driving is another critical domain where foundation agents have major impact opportunities. As a paradigm shift from previous perception-prediction-planning framework, LLMs are exploited to integrate common-sense knowledge and human cognitive abilities into autonomous driving (Wen et al., 2023; Mao et al., 2023), and enable multiple vehicles to realize collaboration over time (Hu et al., 2024). Despite these trends for foundation agents in autonomous control, safety, robustness and reliability of decisions made by agents requires rigorous evaluation and assurance in real-world deployment.

Healthcare. Human-Agent collaborative decision making in healthcare has the opportunity to enhance diagnostic accuracy, streamline operations, and provide personalized treatment options. Current research in foundation agents for healthcare is rare but has shown effectiveness in assessing patient's state in ICU (intensive care unit) rooms (Durante et al., 2024) and designing treatment plans through discussions among LLMs as different medical experts (Tang et al., 2023). However, agents in healthcare still raise high uncertainty and distrust from clinicians. Inspired by drawing biomedical literature evidence via LLM (Yang et al., 2023b) or involving human experts in the intermediate stages of medical decision making (Zhang et al., 2024), real-world healthcare can be realized through consistent collaboration between humans and foundation agents in the future.

Science. Scientific discovery and research can be revolutionized by foundation agents with significantly accelerated data analysis and experimentation processes, leading to faster and more accurate insights. AlphaFold 3 (Abramson et al., 2024) has demonstrated the feasibility of high accuracy modelling across biomolecular space within a single unified model, consistent with the characteristics of foundation agents that have a unified representation and policy interface across state-action spaces. Aligned with LLMs and empowered by external tools, chemistry agents (Boiko et al., 2023; M. Bran et al., 2024) show their potential for accelerating research and discovery. They can be applied in organic synthesis, reaction optimization, drug discovery, and materials design with capabilities of (semi-) autonomous hypothesis generation, experimental design and execution. However, the deployment of foundation agents in scientific research poses potential biases in data interpretation, unintended experimental errors, and ethical concerns regarding the transparency and accountability of findings, requiring rigorous oversight, comprehensive validation and ethical guidelines.

7. Conclusion

We take the position that foundation agents hold the potential to alter the landscape of agent learning for decision making, akin to the revolutionary impact of foundation models in language and vision. We support this position by formulating the notion of foundations agents and specifying their roadmap and key challenges. Demonstrated by recent work and a case study, we show that the enhanced perception, adaptation, and reasoning abilities of agents not only address limitations of conventional RL, but also hold the key to unleash the full potential of foundation agents in real-world decision making. However, due to the intrinsic complexity of decision making, existing work towards foundation agents is still at a starting point. Future directions involve but are not limited to learning a unified representation of diverse environments and tasks, establishing theoretical foundations for a unified policy interface and learning from open-ended tasks at scale. Bridging these gaps will pave the way for more robust and versatile foundation agents, contributing to the evolution of artificial general intelligence.

Acknowledgements

This work is supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27010103) and the Youth Innovation Promotion Association CAS.

Impact Statement

The development and deployment of foundation agents carry profound implications for society, shaping the landscape of AI and decision-making systems. The potential broader impact spans ethical considerations, societal consequences, and the democratization of AI. Ethically, foundation agents could contribute to more transparent and accountable decision-making processes, emphasizing the importance of fairness, interpretability, and adherence to ethical standards. However, it also raises concerns about privacy, bias, and unintended consequences, necessitating robust ethical frameworks and governance. Societally, widespread adoption of foundation agents has the potential to revolutionize various sectors, from autonomous control and healthcare to scientific research and beyond, enhancing efficiency and accessibility. Yet, it requires careful consideration of its impact on employment, socioeconomic disparities, and digital divide issues. Striking a balance between innovation and responsible deployment is crucial to harness the full societal potential of foundation agents while mitigating unintended negative consequences. Future developments in this space should be guided by a commitment to ethical principles, social responsibility, and the promotion of positive societal transformations.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

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

Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., Jaakkola, T., Tenenbaum, J., Kaelbling, L., Srivastava, A., and Agrawal, P. Compositional foundation models for hierarchical planning. *arXiv preprint arXiv:2309.08587*, 2023.

Anthropic. Model card and evaluations for claude models. *https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf*, 2023.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., and Efros, A. A. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.

Bauer, J., Baumli, K., Behbahani, F., Bhoopchand, A., Bradley-Schmieg, N., Chang, M., Clay, N., Collister, A., Dasagi, V., Gonzalez, L., et al. Human-timescale adaptation in an open-ended task space. In *International Conference on Machine Learning*, pp. 1887–1935. PMLR, 2023.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Boige, R., Flet-Berliac, Y., Flajolet, A., Richard, G., and Pierrot, T. Pasta: Pretrained action-state transformer agents. *arXiv preprint arXiv:2307.10936*, 2023.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brandfonbrener, D., Nachum, O., and Bruna, J. Inverse dynamics pretraining learns good representations for multitask imitation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kjMGHTo8Cs.

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.

Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Cai, Y., Zhang, C., Shen, W., Zhang, X., Ruan, W., and Huang, L. Reprem: Representation pre-training with masked model for reinforcement learning. *arXiv preprint arXiv:2303.01668*, 2023.

Carroll, M., Paradise, O., Lin, J., Georgescu, R., Sun, M., Bignell, D., Milani, S., Hofmann, K., Hausknecht, M., Dragan, A., et al. Uni [mask]: Unified inference in sequential decision problems. *Advances in neural information processing systems*, 35:35365–35378, 2022.

Chebotar, Y., Vuong, Q., Hausman, K., Xia, F., Lu, Y., Irpan, A., Kumar, A., Yu, T., Herzog, A., Pertsch, K., et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pp. 3909–3928. PMLR, 2023.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Chen, P.-L. and Chang, C.-S. Interact: Exploring the potentials of chatgpt as a cooperative agent. *arXiv preprint arXiv:2308.01552*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Došilović, F. K., Brčić, M., and Hlupić, N. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215. IEEE, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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. In *International Conference on Machine Learning*, pp. 8469–8488. PMLR, 2023.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J. B., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Durante, Z., Sarkar, B., Gong, R., Taori, R., Noda, Y., Tang, P., Adeli, E., Lakshmikanth, S. K., Schulman, K., Milstein, A., et al. An interactive agent foundation model. *arXiv preprint arXiv:2402.05929*, 2024.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

Forestier, S., Portelas, R., Mollard, Y., and Oudeyer, P.-Y. Intrinsically motivated goal exploration processes with automatic curriculum learning. *Journal of Machine Learning Research*, 23(152):1–41, 2022.

Ge, Y., Hua, W., Ji, J., Tan, J., Xu, S., and Zhang, Y. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023.

Grillotti, L. and Cully, A. Unsupervised behavior discovery with quality-diversity optimization. *IEEE Transactions on Evolutionary Computation*, 26(6):1539–1552, 2022.

Gu, P., Zhao, M., Chen, C., Li, D., Hao, J., and An, B. Learning pseudometric-based action representations for offline reinforcement learning. 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. 7902–7918. PMLR, 17–23 Jul 2022.

Guo, J., Yang, B., Yoo, P., Lin, B. Y., Iwasawa, Y., and Matsuo, Y. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv preprint arXiv:2309.17277*, 2023.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hu, C., Fu, J., Du, C., Luo, S., Zhao, J., and Zhao, H. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

Hu, S., Fang, Z., Fang, Z., Chen, X., and Fang, Y. Agentscodriver: Large language model empowered collaborative driving with lifelong learning. *arXiv preprint arXiv:2404.06345*, 2024.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

Izquierdo, I., Medina, J. H., Vianna, M. R., Izquierdo, L. A., and Barros, D. M. Separate mechanisms for short-and long-term memory. *Behavioural brain research*, 103(1): 1–11, 1999.

Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235313679.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., and Tu, Z. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Laskin, M., Wang, L., Oh, J., Parisotto, E., Spencer, S., Steigerwald, R., Strouse, D., Hansen, S., Filos, A., Brooks, E., et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Lee, K.-H., Nachum, O., Yang, M. S., Lee, L., Freeman, D., Guadarrama, S., Fischer, I., Xu, W., Jang, E., Michalewski, H., et al. Multi-game decision transformers. *Advances in Neural Information Processing Systems*, 35: 27921–27936, 2022.

Lehman, J. and Stanley, K. O. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

Liang, X., Du, X., Wang, G., and Han, Z. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology*, 68(2):1243–1253, 2019.

Liu, B., Mazumder, S., Robertson, E., and Grigsby, S. Ai autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine*, 44(2):185–199, 2023a.

Liu, F., Liu, H., Grover, A., and Abbeel, P. Masked autoencoding for scalable and generalizable decision making. *Advances in Neural Information Processing Systems*, 35: 12608–12618, 2022.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.

Ma, Y. J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., and Zhang, A. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

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: Arxiv-2310.12931*, 2023.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., and Fox, D. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Mao, J., Ye, J., Qian, Y., Pavone, M., and Wang, Y. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023.

Meng, L., Wen, M., Yang, Y., Le, C., Li, X., Zhang, W., Wen, Y., Zhang, H., Wang, J., and Xu, B. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all smac tasks. *arXiv preprint arXiv:2112.02845*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Nakabi, T. A. and Toivanen, P. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy, Grids and Networks*, 25: 100413, 2021.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Video generation models as world simulators. *https://openai.com/research/video-generation-models-as-world-simulators*, 2024.

Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degrave, J., Piot, B., Perolat, J., et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Pinto, V., Gamage, C., Stephenson, M., and Renz, J. The difficulty of novelty detection and adaptation in physical environments. In *Australasian Joint Conference on Artificial Intelligence*, pp. 28–40. Springer, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Sang, T., Tang, H., Ma, Y., Hao, J., Zheng, Y., Meng, Z., Li, B., and Wang, Z. Pandr: Fast adaptation to new environments from offline experiences via decoupling policy and environment representations. In *International Joint Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:247996930.

Schwarzer, M., Rajkumar, N., Noukhovitch, M., Anand, A., Charlin, L., Hjelm, R. D., Bachman, P., and Courville, A. C. Pretraining representations for data-efficient reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12686–12699, 2021.

Schwarzer, M., Ceron, J. S. O., Courville, A., Bellemare, M. G., Agarwal, R., and Castro, P. S. Bigger, better, faster: Human-level atari with human-level efficiency. In *International Conference on Machine Learning*, pp. 30365–30380. PMLR, 2023.

Shafiullah, N. M., Cui, Z., Altanzaya, A. A., and Pinto, L. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35: 22955–22968, 2022.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

Shi, L. X., Hu, Z., Zhao, T. Z., Sharma, A., Pertsch, K., Luo, J., Levine, S., and Finn, C. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910*, 2024.

Shinn, N., Labash, B., and Gopinath, A. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pp. 9870–9879. PMLR, 2021.

Sun, Y., Ma, S., Madaan, R., Bonatti, R., Huang, F., and Kapoor, A. Smart: Self-supervised multi-task pretraining with control transformers. *arXiv preprint arXiv:2301.09816*, 2023.

Tang, H., Meng, Z., Hao, J., Chen, C., Graves, D., Li, D., Yu, C., Mao, H., Liu, W., Yang, Y., et al. What about inputting policy in value function: Policy representation and policy-extended value function approximator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8441–8449, 2022.

Tang, X., Zou, A., Zhang, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Team, O. E. L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Todorov, E. General duality between optimal control and estimation. In *2008 47th IEEE conference on decision and control*, pp. 4286–4292. IEEE, 2008.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning*, pp. 3540–3549. PMLR, 2017.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

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*, 2023a.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024.

Wang, W. W., Han, D., Luo, X., Shen, Y., Ling, C., Wang, B., and Li, D. Toward open-ended embodied tasks solving. *arXiv preprint arXiv:2312.05822*, 2023b.

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*, 2023c.

Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., and Liang, Y. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023d.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Wei, Y., Sun, Y., Zheng, R., Vemprala, S., Bonatti, R., Chen, S., Madaan, R., Ba, Z., Kapoor, A., and Ma, S. Is imitation all you need? generalized decision-making with dual-phase training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16221–16231, 2023.

Wen, L., Fu, D., Li, X., Cai, X., Ma, T., Cai, P., Dou, M., Shi, B., He, L., and Qiao, Y. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023.

Wu, P., Majumdar, A., Stone, K., Lin, Y., Mordatch, I., Abbeel, P., and Rajeswaran, A. Masked trajectory models for prediction, representation, and control. *arXiv preprint arXiv:2305.02968*, 2023.

Xie, Z., Lin, Z., Ye, D., Fu, Q., Wei, Y., and Li, S. Future-conditioned unsupervised pretraining for decision transformer. In *International Conference on Machine Learning*, pp. 38187–38203. PMLR, 2023.

Xu, M., Shen, Y., Zhang, S., Lu, Y., Zhao, D., Tenenbaum, J., and Gan, C. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pp. 24631–24645. PMLR, 2022.

Yang, M. and Nachum, O. Representation matters: Offline pretraining for sequential decision making. In *International Conference on Machine Learning*, pp. 11784–11794. PMLR, 2021.

Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., and Abbeel, P. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023a.

Yang, Q., Hao, Y., Quan, K., Yang, S., Zhao, Y., Kuleshov, V., and Wang, F. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023b.

Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., and Schuurmans, D. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

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

Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., and Gan, C. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023a.

Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L. M., Caterino, J., Zhang, P., et al. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Zhou, Z., Kearnes, S., Li, L., Zare, R. N., and Riley, P. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.

Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

A. Distinctions between Foundation Agents, Task-Specific Agents and LLM-based Agents

Compared to conventional task-specific agents developed by RL or IL approaches, the foundation agent has the capabilities of processing multi-modalities, adapting to new tasks and domains, and few- or zero-shot generalization with prompting. Compared to LLM-based agents that mainly carry out linguistic or knowledge-based tasks, the foundation agent does physical planning, optimizes decisions, and plays two- or multi-player games in physical and virtual worlds. An LLM may be part of a foundation agent to enable the agent to align with the world model and human society created by languages. However, as stated in Section 6.1, the morphology of a foundation agent is not necessarily restricted to a composition of several existing foundation models. Instead, an LLM-based agent in current research (Wang et al., 2024) places LLMs at the core of decision making.

B. Computational Requirements of Potentially Foundation Agents

Table 2 shows the model size and computational requirements of some generalist agents and our case study towards foundation agents. Recent works (Lee et al., 2022; Brohan et al., 2022; 2023; Wei et al., 2023; Chebotar et al., 2023; Durante et al., 2024) demonstrate that these agents are able to continue to scale to larger dataset sizes and model sizes with performance improvement. Therefore, training a foundation agent may require much more computational resources than expected, and scaling issues involving hardware acceleration, parallel processing, and model complexity and optimization need to be considered. In future work, the scaling law concerning the data and model size for training foundation agents needs to be addressed.

Table 2. Model size and computational requirements of current potential foundation agents and our case study. Results except for the last row are from original papers. Note that these agents are still far from the truly foundation agent as defined in our paper, and we only include those whose computational requirements are explicitly described in their original papers.

Model	Params	Computational Requirements
Gato (Reed et al., 2022)	1.18B	Training 4 days on 256 TPUv3
Multi-Game Decision Transformer (Lee et al., 2022)	200M	Preraining 8 days on 64 TPUv4
DualMind (Wei et al., 2023)	175M (pretraining) 51.1M (IL from prompts)	40h in Phase I on 40 V100 12h in Phase II on 16 V100
Interactive Agent Foundation Model (Durante et al., 2024)	277M	Pretraining 175h on 16 V100
Our case	25M	Pretraining 8h on 4 RTX4090

C. Case Study in DMC and Atari

C.1. Environments and Datasets

DeepMind Control Suite. The DeepMind Control Suite (Tassa et al., 2018) is a set of physics-based simulation environments, containing continuous control tasks with image observations spanning diverse domains and tasks. Different domains are associated with different state-actions spaces and each task is associated with a particular MDP within each domain. We select 5 tasks for pretraining and one unseen task for fine-tuning as shown in Table 3. For each task, we collect trajectories using pixels from the training progress of a SAC (Haarnoja et al., 2018) agent for 1M steps to include both expert and non-expert data for pretraining. For fine-tuning, we randomly sample 10% trajectories from the full replay buffer of the SAC agent with diverse return distribution.

Table 3. Domains and tasks used in pretraining and fine-tuning on DMC.

Phase	Domain	Task
Pretraining	cartpole	swingup
	hopper	hop
	cheetah	run
	walker	stand
	walker	run
Fine-tuning	walker	walk

Atari. The Atari benchmark (Bellemare et al., 2013) is a collection of video game environments with distinctly different dynamics, rewards, and agent embodiments. We select 5 games for pretraining and one unseen game for fine-tuning as shown in Table 4. For each game, we collect trajectories using pixels from the training progress of a DQN (Mnih et al., 2015) agent for 1M steps to include both expert and non-expert data for pretraining. For fine-tuning, we randomly sample 10% trajectories from the full replay buffer of the DQN agent with diverse return distribution.

Table 4. Atari games used in pretraining and fine-tuning.

Phase	Game
Pretraining	Asterix
	Breakout
	Gopher
	Qbert
	Seaquest
Fine-tuning	Pong

C.2. Implementation Details

Model Architecture. The implementation of our agent is based on GPT-2 (Brown et al., 2020) with configurations summarized in Table 5. The input for pretraining consists of a sequence of image observations and actions, while during fine-tuning returns are also added as input for RL. The observation tokenizer is a single ResNet (He et al., 2016) block to obtain image patch token embeddings following (Reed et al., 2022). Action tokenizer and return tokenizer are single-layer linear prediction heads. We follow the tokenization scheme and add patch position encodings and local observation position encodings to token embeddings as introduced in (Reed et al., 2022). Similar to (Reed et al., 2022; Lee et al., 2022), we model the data autoregressively using a sequential causal attention masking but allow observation tokens within the same timestep to access each other.

Table 5. Transformer hyperparameters of our agent.

Hyperparameter	Model
Transformer blocks	8
Attention heads	16
Layer width	512
Feedforward hidden size	2048
Shared embedding	True
Layer normalization	Pre-norm
Activation function	GeGLU

Pretraining and Fine-tuning. We jointly pretrain our agent on 5 DMC tasks and 5 Atari games using the pretraining objective of next action prediction $-logP_\theta(a_t|\tau_{0:t-1}, o_t)$ via standard mean-squared error loss for continuous actions or cross-entropy loss for discrete actions. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e-8$ for 1M steps. For pretraining, the learning rate is originally set to be 1e-5 with linear warm-up and cosine schedule decay. We use batch size of 512 and weight decay of 0.1 for pretraining. For fine-tuning, the learning rate keeps constant of 1e-4 with batch size 256 and dropout rate 0.1. During fine-tuning, we model and predict returns and rewards in addition to actions using the learning objective $-logP_\phi(R_t, a_t, r_t|\tau_{0:t-1}, o_t)$, and perform expert action inference as introduced in (Lee et al., 2022). We use context length $L = 50$ for both pretraining and fine-tuning[2]. The size of our pretrained model and corresponding computational requirements can be referred to Table 2. Since this is simply a case study of potential foundation agent in applications of robotics and game play, we only provide the visualization of fine-tuning performance in Figure 2 to demonstrate its feasibility and effectiveness in decision making scenarios. We do not report other experimental results and compare with baselines as done in original research. In addition, the case study only involves the first two steps in the roadmap of foundation agents as specified in Figure 1. We consider aligning the pretrained model in this case study with LLMs in future work.

[2]Our code is based on `https://github.com/microsoft/smart`.