Foundation Models for Decision Making: Problems, Methods, and Opportunities

Sherry Yang*[1,2] Ofir Nachum[1] Yilun Du[3] Jason Wei[1] Pieter Abbeel[2] Dale Schuurmans[1,4]

[1]Google Research, Brain Team, [2]UC Berkeley, [3]MIT, [4]University of Alberta

Foundation models pretrained on diverse data at scale have demonstrated extraordinary capabilities in a wide range of vision and language tasks. When such models are deployed in real world environments, they inevitably interface with other entities and agents. For example, language models are often used to interact with human beings through dialogue, and visual perception models are used to autonomously navigate neighborhood streets. In response to these developments, new paradigms are emerging for training foundation models to interact with other agents and perform long-term reasoning. These paradigms leverage the existence of ever-larger datasets curated for multimodal, multitask, and generalist interaction. Research at the intersection of foundation models and decision making holds tremendous promise for creating powerful new systems that can interact effectively across a diverse range of applications such as dialogue, autonomous driving, healthcare, education, and robotics. In this manuscript, we examine the scope of foundation models for decision making, and provide conceptual tools and technical background for understanding the problem space and exploring new research directions. We review recent approaches that ground foundation models in practical decision making applications through a variety of methods such as prompting, conditional generative modeling, planning, optimal control, and reinforcement learning, and discuss common challenges and open problems in the field.

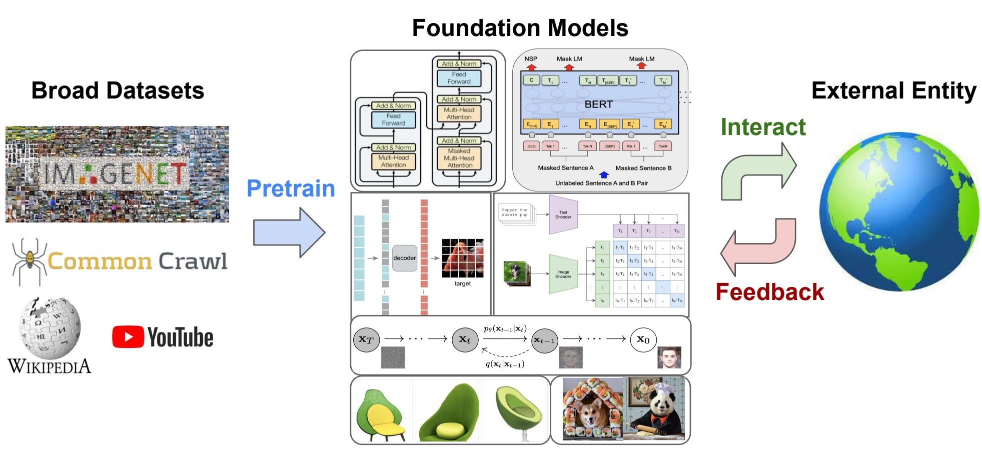

Fig. 1. Overview of foundation models for decision making. Foundation models pretrained on broad data are adapted to accomplish specific tasks by interacting with external entities and receiving feedback.

*Corresponding author: sherryy@berkeley.edu

Contents

1 INTRODUCTION

Foundation models pretrained on broad datasets via self-supervised learning have demonstrated exceptional abilities in knowledge transfer to diverse downstream tasks [Bommasani et al. 2021]. As such models continue to be applied to more complex problems that involve long-term reasoning [Wei et al. 2022a], control [Brohan et al. 2022], search [Strohman et al. 2005], and planning [Huang et al. 2022b], or are deployed in applications such as dialogue, autonomous driving, healthcare, and robotics, they are expected to interface with external entities and agents. For example, in dialogue a language model converses with a human over multiple turns; in robotics a perception-control model executes actions in a real-world environment. These scenarios present new challenges for foundation models, including (1) how to learn from feedback given by an external entity (e.g., human rating of conversation quality), (2) how to adapt to modalities not commonly covered by large language or vision datasets (e.g., robot actions), and (3) how to perform long-term reasoning and planning over the future.

Such questions have traditionally been at the core of sequential decision making [Sutton and Barto 2018], encompassing areas such as reinforcement learning, imitation learning, planning, search, and optimal control. Contrary to the paradigm of foundation models, where broad datasets with billions of images and text tokens are used during pretraining, prior work on sequential decision making has largely focused on task-specific or *tabula rasa* settings with limited prior knowledge [Silver et al. 2017]. Despite a seemingly disadvantageous setup, research in sequential decision making has achieved significant progress in surpassing human performance on tasks such as playing board games [Tesauro 1994] and Atari video games [Mnih et al. 2013], as well as operating robots to complete navigation [Pomerleau 1988] and manipulation tasks [Kalashnikov et al. 2018; Akkaya et al. 2019]. Nevertheless, since these methods learn to solve a task from scratch without broad knowledge from vision, language, or other datasets, they generally struggle with generalization and sample efficiency, e.g., requiring 7 GPU days of interactive game-play to solve a single Atari game [Agarwal et al. 2022]. Intuitively, broad datasets similar to those used for foundation models should also be beneficial for sequential decision making models. For example, there are countless articles and videos on the Internet about how to play Atari games. Similarly, there is a wealth of knowledge about properties of objects and scenes that would be useful to a robot, or about human wants and emotions that could improve a dialogue model.

While research on foundation models and sequential decision making has largely been disjoint due to distinct applications and foci, there is increasing activity at the intersection of these communities. On the foundation models side, with the discovery of emergent properties of large language models, target applications have graduated from simple zero or few-shot vision and language tasks to problems that now involve long-term reasoning [Srivastava et al. 2022; Wei et al. 2022b; Lewkowycz et al. 2022] or multiple interactions [OpenAI 2022]. Conversely, in the sequential decision making communities, researchers inspired by the success of large scale vision and language models have begun to curate ever-larger datasets for learning multimodel, multitask, and generalist interactive agents [Agarwal et al. 2020b; Szot et al. 2021; Fan et al. 2022; Brohan et al. 2022; Reed et al. 2022; Lee et al. 2022]. Further blurring the lines between the two fields, some recent work has investigated the use of pretrained foundation models such as CLIP [Radford et al. 2021] and ViT [Dosovitskiy et al. 2020] to bootstrap the training of interactive agents for visual environments [Khandelwal et al. 2022; Tao et al. 2022], while other work has investigated foundation models as dialogue agents optimized by reinforcement learning with human feedback [Ouyang et al. 2022], and other work has adapted large language models to interact with external tools such as search engines [Komeili et al. 2021; Thoppilan et al. 2022; Lazaridou et al. 2022; Shuster et al.

2022; Yao et al. 2022], calculators [Cobbe et al. 2021; Thoppilan et al. 2022], translators [Thoppilan et al. 2022], MuJoCo simulators [Liu et al. 2022d], and program interpreters [Gao et al. 2022].

Our premise in this report is that research on foundation models and interactive decision making can be mutually beneficial if considered jointly. On one hand, adaptation of foundation models to tasks that involve external entities can benefit from incorporating feedback interactively and performing long-term planning. On the other hand, sequential decision making can leverage world knowledge from foundation models to solve tasks faster and generalize better. With the aim of spurring further research at the intersection of these two fields, we scope the problem space of *foundation models for decision making*. We provide technical tools for understanding current research in the space, review remaining challenges and open problems, and speculate on potential solutions and promising approaches to overcome these challenges.

1.1 Structure of This Report

This report is divided into 5 major sections. In Section 2, we review the relevant background and notations of sequential decision making, and present a few example scenarios where foundation models and decision making are better considered jointly. The subsequent three sections are organized around how foundation models can characterize different components of a decision making system. In Section 3, we discuss how foundation models can serve as generative models of *behavior* (e.g., skill discovery) and generative models of the *environment* (e.g., for conducting model-based rollouts). In Section 4, we discuss how foundation models can serve as representation learners of states, actions, rewards, and transition dynamics (e.g., plug-and-play vision-language models, model-based representation learning). In Section 5, we discuss how language foundation models can serve as interactive agents and environments, enabling new problems and applications to be considered under a sequential decision making framework (language model reasoning, dialogue, tool use). Finally in Section 6, we outline open problems and challenges, and propose potential solutions (e.g., how to leverage broad data, how to structure environments, and what aspects of foundation models and decision making can be improved).

2 PRELIMINARIES

In this section, we review relevant background on sequential decision making, and present example scenarios to illustrate when and why it is better to consider foundation models and decision making jointly.

2.1 Sequential Decision Making Preliminaries

Unlike vision and language domains, where a foundation model is usually trained (and adapted) only once, sequential decision making focuses on learning from *interactive* experience. We outline this formalism and introduce common algorithms for sequential decision making.

2.1.1 Interacting with an Environment.

Sequential decision making problems are most often formalized in terms of a Markov decision process (MDP) [Puterman 1994], which is defined as a tuple $\mathcal{M} := \langle S, A, \mathcal{R}, \mathcal{T}, \mu, \gamma \rangle$ consisting of a state space S, an action space A, a reward function $\mathcal{R} : S \times A \to \Delta(\mathbb{R})$,[†] a transition function $\mathcal{T} : S \times A \to \Delta(S)$, an initial state distribution $\mu \in \Delta(S)$, and a discount factor $\gamma \in [0, 1)$. A policy $\pi : S \to \Delta(A)$ interacts with the environment starting at an initial state $s_0 \sim \mu$. At each timestep

[†] $\Delta(\mathcal{X})$ denotes the simplex over a set $\mathcal{X}$.

$t \geq 0$, an action $a_t \sim \pi(s_t)$ is sampled and applied to the environment, after which the environment transitions into the next state $s_{t+1} \sim \mathcal{T}(s_t, a_t)$ while producing a scalar reward $r_t \sim \mathcal{R}(s_t, a_t)$.[‡]

After π interacts with $\mathcal{M}$ for H timesteps (H can be infinite), an episode (trajectory) is produced $\tau := \{(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)\}$. We use τ_t to denote the tuple (s_t, a_t, r_t), $\tau_{<t}$ to denote a sub-episode up to timestep t, $\tau_{\geq t}$ to denote a sub-episode starting from timestep t and ending at H, $\tau_{t:t+h}$ to denote a sub-episode from timestep t to $t+h$, and τ_s or τ_a to denote only the state or action portion of a trajectory. The return associated with episode τ is defined as the total discounted sum of rewards $R(\tau) := \sum_{t=0}^{H} \gamma^t r_t$. The *trajectory distribution* of a policy $p_\pi(\tau)$ is determined by

$$p_\pi(\tau) = \mu(s_0)\Pi_{t=0}^{H} \pi(a_t|s_t)\mathcal{R}(s_t, a_t)\mathcal{T}(s_{t+1}|s_t, a_t). \tag{1}$$

Trajectories generated by one or multiple policies can be collected in an offline dataset $\mathcal{D}_{\text{RL}} = \{\tau\}$. We distinguish $\mathcal{D}_{\text{RL}}$ from a typical vision or language dataset $\mathcal{D}$; $\tau \sim \mathcal{D}_{\text{RL}}$ is an *interactive* trajectory involving actions and rewards whereas $x \sim \mathcal{D}$ is a *static* image or a text sequence. Nevertheless, foundation model techniques developed for $\mathcal{D}$ can also be apply to $\mathcal{D}_{\text{RL}}$.

2.1.2 *Imitation Learning.*

In standard imitation learning, $\mathcal{R}$, $\mathcal{T}$, and μ are unknown to the agent. Learning solely takes place from a fixed dataset of demonstrations $\mathcal{D}^*_{\text{RL}} = \{(s, a)\}$ previously collected by an expert policy π^* interacting with $\mathcal{M}$ through $a \sim \pi^*(s)$. The goal of imitation learning is to train π on $\mathcal{D}^*_{\text{RL}}$ so that π closely approximates π^* according to some metric, such as the Kullback–Leibler (KL) divergence between the trajectory distributions $D_{\text{KL}}(p_{\pi^*}(\tau)\|p_\pi(\tau))$.

Behavioral cloning (BC). Learning from expert demonstrations leads to the common framing of imitation learning as supervised learning of state to action mappings. Under this framing, behavioral cloning (BC) [Pomerleau 1989] proposes to learn π by minimizing

$$\mathcal{L}_{\text{BC}}(\pi) := \mathbb{E}_{(s,a)\sim\mathcal{D}^*_{\text{RL}}}[-\log \pi(a|s)]. \tag{2}$$

Equation 2 can be viewed as the classification loss (discrete actions) or regression loss (continuous actions) of state to action mappings, connecting BC to supervised learning in vision and language.

2.1.3 *Reinforcement Learning.*

Standard reinforcement learning [Sutton and Barto 2018] aims to maximize the expected returns of a policy through trial-and-error interaction with the environment:

$$J(\pi) := \mathbb{E}\left[\, \textstyle\sum_{t=0}^{H} \gamma^t r_t \middle| \pi, \mathcal{M}\right]. \tag{3}$$

Policy-based methods. One conceptually straightforward way to optimize Equation 3 is through policy gradient, which estimates the gradient of Equation 3 with respect to the policy π, and maximizes $J(\pi)$ directly via gradient ascent. The most commonly used gradient estimator has the form

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\, \textstyle\sum_{t=0}^{H} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}(s_t, a_t)\right], \tag{4}$$

where $\hat{A}$ is some advantage function that can be separately estimated via Monte-Carlo returns from $p_\pi(\tau)$ [Williams 1992]. The biggest drawback of policy gradient is sample inefficiency: since policy gradients are estimated from rollouts, the variance of the gradient estimate is often extreme. To mitigate high variance, various works such as PPO [Schulman et al. 2017] have proposed to improve policy updates through the use of appropriate geometry [Kakade 2001; Peters et al. 2010;

[‡]We will focus on fully observable MDPs in this article, though an MDP can be extended to a partially observable MDP (POMDP) by introducing an observation space $\mathcal{O}$, an emission function $\mathcal{E} : S \to \mathcal{O}$, and the restriction that policies can only depend on observations and previous actions.

Schulman et al. 2015a] or through training a separate critic network to estimate $\hat{A}$ to futher reduce variance at the cost of introducing bias [Sutton et al. 1999; Silver et al. 2014; Schulman et al. 2015b].

Value-based methods. Another family of reinforcement learning methods for optimizing Equation 3, such as Q-learning [Watkins and Dayan 1992], involves learning the optimal value function $Q^*(s_t, a_t)$ by satisfying a set of Bellman *optimality* constraints:

$$Q^*(s_t, a_t) = r_t + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t,a_t)} \left[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \right], \tag{5}$$

after which an optimal policy can be extracted via $\pi^*(\cdot|s_t) = \arg_a \max Q^*(s_t, a)$. Value-based methods are typically more sample efficient than policy-based methods [Gu et al. 2016], but tend to be unstable under function approximation [Sutton and Barto 2018]. At the intersection of policy and value based methods, Actor-Critic methods [Sutton et al. 1999] first learn $Q^\pi(s_t, a_t)$ by satisfying the set of Bellman *expectation* constraints:

$$Q^\pi(s_t, a_t) = r_t + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t,a_t), a_{t+1} \sim \pi(s_{t+1})} \left[Q^\pi(s_{t+1}, a_{t+1}) \right], \tag{6}$$

then plug $\hat{A}(s_t, a_t) = Q^\pi(s_t, a_t)$ into the policy gradient objective, Equation 4, to update the policy. The intuition that the resulting policy learning will be both stable and sample efficient.

Off-policy and offline RL. To further improve the sample efficiency of on-policy methods, a set of off-policy approaches have been proposed for both policy and value based RL [Lillicrap et al. 2015; Mnih et al. 2016; Nachum et al. 2017], where data from sources other than the current policy can be utilized for learning in conjunction with environment interaction. Offline RL [Levine et al. 2020] further considers the setting where an agent only has access to a fixed dataset of previous interactions $\mathcal{D}_{\text{RL}}$, and no further environment access to $\mathcal{T}$ or $\mathcal{R}$ is available. To ensure the learned policy avoids out-of-distribution states and actions, offline RL methods often impose regularization via a divergence between the learned policy and the offline dataset [Wu et al. 2019] or on the learned value function [Kumar et al. 2020]. More recently, some works have explored using additional online access as a finetuning step after offline RL to improve sample efficiency [Nair et al. 2020; Xie et al. 2021; Ball et al. 2023].

Using foundation models for decision making differs from traditional offline RL (with or without online finetuning) in that the latter focuses on learning RL algorithms from task-specific RL datasets $\mathcal{D}_{\text{RL}}$ (i.e., datasets with task-specific states, actions, and rewards), whereas the former focuses on self-supervised learning on diverse data (e.g., data from vision and language domains) followed by task-specific adaptation.

2.1.4 *Planning, Search, and Optimal Control.*

Unlike the model-free RL algorithms outlined above, a broader set of approaches to sequential decision making (e.g., planning, search, optimization-based control, model-based RL) leverage explicit models of the environment. When the true environment dynamics are known (e.g., the rules of a Chess game) and simulation is cheap, planning and search algorithms, such as MCTS [Kocsis et al. 2006] that leverage an accurate simulator, can be highly effective [Silver et al. 2016]. When the environment can be characterized by precise dynamics, such as the constrained movements of a robot arm, approaches in optimal control—such as trajectory optimization [Von Stryk and Bulirsch 1992], shooting [Bock and Plitt 1984], collocation [Von Stryk 1993], and model predictive control [Camacho and Alba 2013]—have long been studied prior to the recent advances in deep learning. In deterministic scenarios, given an environment governed by a known dynamics function

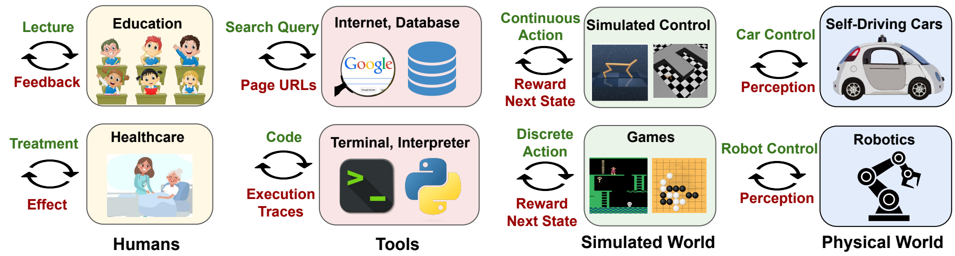

Fig. 2. Example scenarios of adapting foundation models to perform decision making tasks such as interacting with humans, tools, and the simulated and physical world. Actions generated by foundation models and feedback provided by the external entities often reoccur repeatedly in a loop.

$s_{t+1} = f(s_t, a_t)$, optimizing a sequence of actions $a_{0:T}$ to execute in the environment corresponds to

$$a_{0:T} = \arg\max_{a_{0:T}} J(s_0, a_{0:T}) = \arg\max_{a_{0:T}} \sum_{t=0}^{T} r(s_t, a_t) \text{ subject to } s_{t+1} = f(s_t, a_t). \tag{7}$$

Model-based RL [Doya et al. 2002] considers the setting where the environment dynamics are unknown and have to be estimated from samples, after which techniques from search, planning, and optimal control [Doya et al. 2002; Deisenroth and Rasmussen 2011; Tassa et al. 2012; Nagabandi et al. 2018; Kaiser et al. 2019] can be effectively applied given the learned dynamics model.

2.2 Example Scenarios

Before diving into the details of foundation models for decision making, we first discuss a few example scenarios where joint consideration of foundation models and decision making can be highly beneficial. Figure 2 illustrates additional examples where foundation models can interact with external entities (e.g., humans, tools, and simulated and physical worlds).

Learning dialogue agents with human feedback. There has been an increasing demand for large language models to produce likable, factual, and grounded responses to human inquires. With a moderate amount of human feedback, via prompting or reward-based finetuning, langauge models have been able to perform increasingly more complex reasoning and dialogue tasks. Such feedback can be seen as the result of langauge model agents interacting with the external world (i.e., humans). Learning from interaction lies at the center of decision making, and reinforcement learning techniques such as policy gradient introduced in Section 2.1.3 have contributed significantly to the advances of dialogue systems [Ouyang et al. 2022].

The Internet as an environment. While RL with human feedback has demonstrated tremendous empirical success in dialogue [Thoppilan et al. 2022; OpenAI 2022], humans are by no means the only external entity that can provide feedback to improve foundation models through repeated interaction. For instance, the Internet can be viewed as an unbounded environment where an ideal policy should be able to identify the best queries and navigation steps to retrieve optimal answers in a minimal number of interactive steps. Since the Internet is both rich in information and cheap to interact with, it provides a compelling environment to explore decision making techniques. Foundation models are necessary for Internet-scale decision making, as interaction needs to be initiated in a reasonable way to ensure meaningful feedback is obtained for further learning.

Video generation as a universal policy. A central difficulty in learning general-purpose robot agents is the incongruity between the state and action spaces of different environments. This implies that, for example, data collected by different robots cutting an apple or videos of a human cutting an apple cannot be easily combined to train a generalist robot policy, despite the fact that the notions of "cutting" and "apple" are common between these scenarios. With ever-larger text-to-video foundation models being trained on Internet-scale data [Ho et al. 2022; Villegas et al. 2022], it is now possible to recast the problem of policy learning as a text-conditioned video generation problem, where the generation process encompasses both environment modeling and planning. Such a policy-as-video formulation allows a unified interface (i.e., images) for learning and generalization from broad data sources, environments, and tasks.

3 FOUNDATION MODELS AS CONDITIONAL GENERATIVE MODELS

We now examine the first concrete use case of foundation models in decision making: probabilistic modeling of the trajectory distribution $p(\tau)$ from an interactive dataset $\tau \sim \mathcal{D}_{\text{RL}}$. Depending on what part of τ is being modeled, foundation models can serve as conditional generative models of behaviors (i.e. actions) or the underlying world models (i.e., environment dynamics). Below, we first review different generative models and then discuss and explore how they can be used to represent behaviors and models of the environment.

3.1 Generative Model Preliminaries

Many foundation models can be characterized as modeling a (conditional) density $p(x)$ on a large dataset of images or texts $x \sim \mathcal{D}$. For example, x could be an image, a sequence of images, or a sequence of text tokens. Different foundation models differ in their factorizations of $p(x)$. Below, we provide a brief overview of several generative models and their factorizations of $p(x)$.

3.1.1 *Latent Variable Models.*

Latent variable models factorize the unknown data distribution of interest $p(x)$ into a latent variable distribution and a conditional distribution:

$$p(x) = \int p(z)p(x|z)dz, \tag{8}$$

where the latent variable z can be both discrete or continuous. For the special cases when z is discrete and the sum is tractable, or z is continuous and the integral is tractable, one can simply calculate $p(x)$ in closed form to support efficient maximum likelihood estimation on a given dataset. However, for the more general cases when the requisite sum or integral is intractable, techniques like VAEs [Kingma and Welling 2013] are applied to optimize the evidence lower-bound (ELBO) of $p(x)$ using a variational posterior $q(z|x)$:

$$\mathcal{L}_{\text{VAE}}(p, q) = \mathbb{E}_{x\sim\mathcal{D}, z\sim q(z|x)}\left[-\log p(x|z)\right] + \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}\left(q(z|x)\|p(z)\right)\right]. \tag{9}$$

As an extension of VAE, VQ-VAE [Van Den Oord et al. 2017] uses a codebook to discretize the continuous latent representation to learn a more compact, discrete representation of the data.

3.1.2 *Autoregressive Sequence Models.*

Autoregressive sequence models have been popularized by transformer-based language models [Vaswani et al. 2017; Brown et al. 2020]. At their core, autoregressive models factorize any joint distribution over a sequence $x = (x_1, ...x_L)$ in an autoregressive manner:

$$p(x) = \Pi_{\ell=1}^{L} p(x_\ell | x_{<\ell}). \tag{10}$$

Under this factorization, estimating the density $p(x)$ reduces to learning each conditional factor $p(x_\ell|x_{<\ell})$ which can be parametrized by a transformer.

$$\mathcal{L}_{\text{LM}}(p) = \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{\ell=1}^{L} -\log p(x_\ell|x_{<\ell})\right]. \tag{11}$$

3.1.3 *Diffusion Models.*

Diffusion models [Sohl-Dickstein et al. 2015; Ho et al. 2020; Kingma et al. 2021] are a class of latent variable models that factorize the data distribution $p(x)$ as a Markov chain of Gaussian transitions from a noise distribution of the same dimension:

$$p(x) = \int p(x_K)\Pi_{k=1}^{K} p(x_{k-1}|x_k)dx_{1:K}, \tag{12}$$

where $p(x_K) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(x_{k-1}|x_k) := \mathcal{N}(\mu(x_k, k), \sigma(x_k, k))$. The forward diffusion process corrupts x by iteratively adding Gaussian noise with a fixed variance schedule. The reverse process then achieves data generation by approximating the noise that corrupted x during the forward process.

3.1.4 *Energy-Based Models.*

Energy-based models [LeCun et al. 2006; Du and Mordatch 2019] are a class of models that represent data distributions $p(x)$ by an unnormalized distribution parameterized by a learned energy function:

$$p(x) = \frac{e^{-E(x)}}{Z}, \tag{13}$$

where E is the energy function and $Z = \int e^{-E(x)}dx$ is the partition function. To sample from the underlying distribution $p(x)$, one typically runs an MCMC procedure such as Langevin dynamics to sample from the underlying distribution.

3.2 Generative Models of Behavior

The generative models introduced above have mostly been applied to text or image data $x \sim \mathcal{D}$. Decision making, on the other hand, is concerned with task specific *interactive* data $\tau \sim \mathcal{D}_{\text{RL}}$ that distinguishes state, action, and reward labels. We will see how different generative models can be adopted to model agent behaviors (this subsection) and environment dynamics (next subsection), as illustrated in Figure 3.

3.2.1 *Foundation Models as Behavioral Priors.*

When the interactive data $\mathcal{D}_{\text{RL}}$ contains diverse behaviors such as "pick up objects", "move objects horizontally", or "place objects", these behaviors can be composed to complete tasks that were not present in $\mathcal{D}_{\text{RL}}$. Foundation models can be used to model such "behavioral priors" (also known as "skills" or "options"). In this approach, pretraining generally involves maximum likelihood estimation of actions conditioned on some trajectory level information. Different tractable approximations can be leveraged to optimize this underlying training objective. For instance, the VAE objective from Equation 9 can be directly instantiated, where the encoder q takes a trajectory τ or some future goal as input and the decoder π produces the sequence of actions as outputs [Ajay et al. 2020; Lynch et al. 2020]:

$$\mathcal{L}_{\text{VAE}}(\pi, q) = \mathbb{E}_{\tau\sim\mathcal{D}_{\text{RL}}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|s_t, z)\right] + \mathbb{E}_{\tau\sim\mathcal{D}_{\text{RL}}}\left[D_{\text{KL}}(q(z|\tau)\|p(z|s_0))\right]. \tag{14}$$

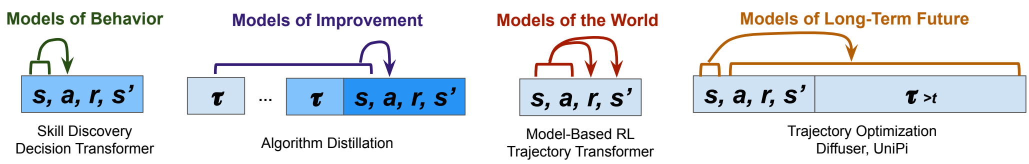

Fig. 3. Illustrations of how conditional generative models can model behaviors, improvements, environments, and long-term futures given a trajectory $\tau \sim \mathcal{D}_{\text{RL}}$. Dark blue indicates transitions with higher rewards. Models of behavior (Decision Transformers [Lee et al. 2022]) and self-improvement (Algorithm Distillation [Laskin et al. 2022]) require near-expert data. Models of the world (Trajectory Transformer [Janner et al. 2021]) and long-term future (UniPi [Du et al. 2023b]) generally require data with good coverage.

The posterior distribution $q(z|\tau)$ can represent a diverse set of behavioral priors when τ is drawn from a wide set of related tasks. Since the posterior depends on future information, the prior $p(z|s_0)$ is usually constrained to only depend on the past so that behaviors can be correctly sampled at test time.

Similarly, the autoregressive sequence modeling objective from Equation 11 can also be instantiated to model behavioral priors [Shafiullah et al. 2022], resulting in a policy that can depend on the history of interaction $\pi(a_t|s_t, \tau_{<t})$. Such dependence is less common in Markovian environments, but has shown empirical benefits [Brohan et al. 2022]. When the dataset consists of expert data $\mathcal{D}^*_{\text{RL}}$, one can learn transformer-based BC policies by optimizing the sequence modeling objective where an autoregressive transformer encodes the history $(\tau_{<t}, s_t)$ and decodes the next action a_t as:

$$\mathcal{L}_{\text{LM}}(\pi) = \mathbb{E}_{\tau \sim \mathcal{D}^*_{\text{RL}}}[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{<t}, s_t)]. \tag{15}$$

An additional conditioning variable z that captures trajectory-level information such as the goal or return $z(\tau) = R(\tau)$ has been introduced in goal or return conditioned supervised learning [Schmidhuber 2019; Kumar et al. 2019; Brandfonbrener et al. 2022; Paster et al. 2022; Yang et al. 2022b]:

$$\mathcal{L}_{\text{LM}}(\pi) = \mathbb{E}_{\tau \sim \mathcal{D}_{\text{RL}}}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{<t}, s_t, z(\tau))\right]. \tag{16}$$

When behavior generation is conditioned on high returns, intuitively, desirable behavior is encouraged [Chen et al. 2021].

One can also utilize a diffusion model to model the conditional distribution of behaviors [Ajay et al. 2022] by maximizing the likelihood in Equation 12:

$$\mathcal{L}_{\text{Diffusion}}(\pi) = \mathbb{E}_{\tau \sim \mathcal{D}_{\text{RL}}, k \sim K}\left[\sum_{t=0}^{H} -\log \pi(a_t^{k-1}|a_t^k, s_t, z(\tau))\right]. \tag{17}$$

To extract desirable behavior from a diffusion model when conditioned on high reward, one can sample trajectories with high likelihood by using reward as classifier-free guidance [Ho and Salimans 2022].

Other conditional generative models that use normalizing flows [Singh et al. 2020], generative adversarial networks [Ho and Ermon 2016], and energy-based models [Florence et al. 2022] have also been proposed for modeling behavioral priors from $\mathcal{D}_{\text{RL}}$.

3.2.2 *Generalist Agents Trained on Massive Behavior Datasets.*

A key advantage to generative modeling of behaviors lies in scaling up; despite different tasks possessing different observations and rewards, there are often meaningful behaviors shared across tasks (e.g., "moving left" has similar meaning in navigation, game playing, and robot manipulation tasks). Inspired by the scaling success of transformers, generalist agents modeling sequences of diverse behaviors have been developed for simulated tasks [Shafiullah et al. 2022], over 40 Atari games [Lee et al. 2022], over 700 real-world robot tasks [Brohan et al. 2022], and over 600 distinct tasks with varying modalities, observations and action specifications [Reed et al. 2022]. This has led to generalist agents that are able to play video games, caption images, chat, perform robot tasks, significantly better than specialist agents trained on single tasks. Such works have also demonstrated the benefit of scaling model parameters and the number of training tasks.

While combining multiple task-specific datasets $\mathcal{D}_{\text{RL}}$ into a large multi-task dataset as described above is one way to scale up behavior modeling, exploiting Internet-scale collections of text and video data $\mathcal{D}$ is another viable approach to scaling effectively. Internet-scale text and video data is abundant in quantity but typically has limited action annotations compared to $\mathcal{D}_{\text{RL}}$. Nevertheless, previous work has still incorporated such datasets. For instance, Gato [Reed et al. 2022] approaches this issue with universal tokenization, so that data with and without actions can be jointly trained using large sequence models. UniPi [Du et al. 2023b] directly learns to predict robotic videos and trains a separate inverse model to infer actions from generated videos. Applying inverse dynamics models to label large video data (e.g., from YouTube) is also applicable to other domains such as self-driving cars [Zhang et al. 2022a] and video game playing [Baker et al. 2022; Venuto et al. 2022].

3.2.3 *Large Scale Online Learning.*

An alternative approach to assuming access to large-scale behavior datasets, online access to massive online game simulators has enabled "large-scale" online RL models to be trained in games such as DoTA [Berner et al. 2019] and StarCraft [Vinyals et al. 2019] using policy gradient or actor-critic algorithms. Similarly, domain randomization [Tobin et al. 2017] has been proposed to leverage online access to diverse generated environments to help bridge the sim-to-real gap in robotics. These large scale online training schemes, however, have not been able to leverage foundation models. An important direction for future work is to explore how one can utilize and learn generative models similarly in massive online settings.

3.2.4 *Generative Models of Exploration and Self-Improvement.*

Generative models of behavior can also be extended to model meta-level processes, such as exploration and self-improvement, whenever the dataset itself $\mathcal{D}_{\text{RL}}$ embodies exploratory and self-improving behavior (e.g., the replay buffer of a policy gradient agent trained from scratch) [Laskin et al. 2022]. That is, unlike other meta-RL methods, which usually train in online settings by maximizing multi-episodic value functions [Wang et al. 2016; Duan et al. 2016], algorithm distillation imitates the action sequence of a multi-episodic improvement process from $\mathcal{D}_{\text{RL}}$ by using a transformer-based sequence model inspired by the zero-shot ability of language models, and adapts to downstream tasks purely in-context without updating any network parameters.

Similar to algorithm distillation, which prompts an agent with its prior learning experience, corrective re-prompting also treats long-horizon planning as an in-context learning problem, but uses corrective error information as prompts, essentially incorporating feedback from the environment as an auxiliary input to improve the executability of a derived plan [Raman et al. 2022].

3.3 Generative Models of the World

In addition to learning models of behaviors, generative models can also learn models of the world—i.e., the transition dynamics $\mathcal{T}$ and the reward function $\mathcal{R}$—from the offline dataset $\mathcal{D}_{\text{RL}}$. Conditional generation from a world model is analogous to model-based rollouts, which can be used to improve a policy.

3.3.1 One-Step Prediction of Reward and Dynamics for Model-based Planning.

One can view learning models of $\mathcal{T}$ and $\mathcal{R}$ as a generative modeling problem given trajectories from an offline dataset $\tau \sim \mathcal{D}_{\text{RL}}$. Since $\mathcal{D}_{\text{RL}}$ also contains actions from a behavior policy π, then π, $\mathcal{T}$, and $\mathcal{R}$ can be jointly modeled with a single generative procedure. Specifically, the joint distribution of a trajectory $p(\tau)$ can be factored autoregressively into an environment component and a policy component,

$$p(\tau) = \Pi_{t=0}^{H} p(s_t, r_t, a_t|\tau_{<t}) = \Pi_{t=0}^{H} \mathcal{T}(s_t|\tau_{<t}) \cdot \pi(a_t|\tau_{<t}, s_t) \cdot \mathcal{R}(r_t|\tau_{<t}, s_t, a_t), \tag{18}$$

so that maximum likelihood estimation of $p(\tau)$ using $\mathcal{D}_{\text{RL}}$ under this factorization naturally decomposes into learning the environment dynamics $\mathcal{T}, \mathcal{R}$ and the policy π that produced the dataset $\mathcal{D}_{\text{RL}}$.

Unlike language models where words exist in a common discrete space, here the states, actions and rewards in τ can all be expressed in different modalities, which poses challenges to sequentially modeling τ. As a workaround, the Trajectory Transformer [Janner et al. 2021] discretizes each dimension of states, actions, and rewards in a continuous control task before applying a GPT-style autoregressive model on the discretized tokens. Discretization is more challenging in image-based domains, where learning a latent representation of an image space and latent dynamics model is more common. Here one can introduce a per-step latent variable z_t into the sequence modeling objective in Equation 18:

$$p(\tau) = \Pi_{t=0}^{H} \int_{z_t} \mathcal{T}_{\text{enc}}(z_t|\tau_{<t}) \cdot \mathcal{T}_{\text{dec}}(s_t|\tau_{<t}, z_t) \cdot \pi(a_t|\tau_{<t}, z_t) \cdot \mathcal{R}(r_t|\tau_{<t}, z_t, a_t) dz_t, \tag{19}$$

where $\mathcal{T}_{\text{enc}}(z_t|\tau_{<t})$ encodes the history into the next step's latent state, $\mathcal{T}_{\text{dec}}(s_t|\tau_{<t}, z_t)$ decodes the next step's observation, and the policy π and reward $\mathcal{R}$ can take latent state z_t as input. Along this line, both Hafner et al. [2020] and Chen et al. [2022b] apply a sequential VAE [Zhu et al. 2020] to optimize the ELBO of Equation 19, and parametrize the latent dynamics model using an RNN or transformer based state space model respectively. Similarly, [Micheli et al. 2022; Ozair et al. 2021; Seo et al. 2022b,a] usesd VQ-VAE or masked autoencoders (MAE) to map image-based observations into discrete tokens before learning a transformer or latent state space dynamics model on the discretized observations.

The various ways a learned world model can be used to infer a high quality policy have been method and task specific. For example, heuristic decoding such as return guided beam search and MCTS have been applied to policy optimization [Janner et al. 2021; Sun et al. 2022; Ozair et al. 2021]. Separate actor and critic pairs have also been trained using rollouts from a latent world model (also referred to as "imagination") without requiring generating image-based observations [Racanière et al. 2017; Hafner et al. 2019]. A world model, when trained to predict observations and actions in the original input space, can also be used to generate additional training data for model-free RL [Sutton 1990; Feinberg et al. 2018; Kaiser et al. 2019; Agarwal et al. 2020a] under the Dyna framework [Sutton and Barto 2018] or to generate additional input context to a policy [Du and Narasimhan 2019].

3.3.2 *Planning with Generative Models of Long-term Future.*

Instead of autoregressively factoring τ by time step as in Equation 18, one can also directly model the joint distribution of τ across all time steps at once using a diffusion model [Du et al. 2019; Janner et al. 2022]:

$$p(\tau) = p(s_0, a_0, r_0, \dots, s_H, a_H, r_H) = \int p(\tau_K)\Pi_{k=1}^{K} p(\tau_{k-1}|\tau_k) d\tau_{1:K}. \quad (20)$$

By learning a trajectory level generative model, planning can be more easily integrated with dynamics modelling by sampling from the composed distribution

$$\tilde{p}(\tau) \propto p(\tau)z(\tau), \quad (21)$$

where $z(\tau)$ specifies the trajectory-level properties that one wishes to control. For instance, Janner et al. [2022] uses trajectory returns as $z(\tau)$ to guide a reverse diffusion process towards sampling high-return trajectories. Ajay et al. [2022] further demonstrate that $z(\tau)$ can represent different trajectory-level properties such as goals, skills, and dynamics constraints, where classifier-free guidance can be applied to conditionally sample trajectories that satisfy the desired properties. Going beyond low dimensional state action spaces, [Du et al. 2023b] also show that diffusion models of long-term futures can also be applied to high-dimensional video data τ, using $z(\tau)$ as text descriptions, effectively improving decision making with large-pretrained text-video foundation models.

In addition to the benefit of flexible conditioning (e.g., on returns, goals, constraints, skills, texts), sampling from the composed distribution in Equation 21 holds the promise of accurate long horizon planning, since sampling an entire trajectory does not suffer from compounding error when rolling out single-step dynamics. Beyond diffusion models, EBMs can also be used to model the joint trajectory distributions $p(\tau)$, including conditioning on latent trajectory properties $z(\tau)$, which might provide a natural approach to satisfying multiple desirable properties, such as high return and safety [Du et al. 2020; Liu et al. 2022b].

4 FOUNDATION MODELS AS REPRESENTATION LEARNERS

In this section, we discuss foundation models for decision making that leverage representation learning for knowledge compression. On one hand, foundation models can extract representations from broad image and text data, $\mathcal{D}$, resulting in a plug-and-play style of knowledge transfer to vision and language based decision making tasks. On the other hand, foundation models can also be used to support task-specific representation learning via task-specific objectives and interactive data, $\mathcal{D}_{\text{RL}}$.

4.1 Plug-and-Play

Off-the-shelf foundation models pretrained on Internet-scale text and image data can be used as preprocessors or initializers for various perceptual components of decision making agents. For instance, when an agent's perception is based on images, contrastive learning [Chen et al. 2020] and masked autoencoding [He et al. 2022] can be directly applied to the agent's image observations, providing state representations that can be further finetuned by BC or RL objectives [Sermanet et al. 2018; Kostrikov et al. 2020; Laskin et al. 2020; Xiao et al. 2022]. When agent actions can be characterized by natural language (e.g., "move to the left then pick up the cup"), pretrained language models can be used to generate higher-level plans for longer-horizon tasks, with the hope that language based descriptions of actions generalize better than low-level motor controls [Huang et al. 2022a; Ahn et al. 2022; Wang et al. 2023; Driess et al. 2023]. When agent observations consist of both images and text descriptions, vision-language captioning models can further enrich agent observations with language descriptions [Tam et al. 2022; Du et al. 2023a; Driess et al. 2023]. Vision-language models such as CLIP and PaLI [Chen et al. 2022a] are further able to provide task feedback and reward information by aligning image and language modalities in the agent's observation and goal space [Huang et al. 2022a; Mahmoudieh et al. 2022; Fan et al. 2022]. Even in the case where an agent's states, actions, and rewards do not consist of images or text, pretrained language models, perhaps surprisingly, have still been found useful as policy initializers for offline RL [Reid et al. 2022], online RL [Li et al. 2022b], and structured prediction tasks [Lu et al. 2021].

Plug-and-play foundation models are generally more natural when the decision making task concerns real-world images or texts. Plug-and-play is less applicable to decision making tasks when there are idiosyncratic, domain specific state action spaces, which we will discuss in Section 4.3. We will further discuss the challenges of bridging general image and text data with task-specific decision making data in Section 6.1.

4.2 Vision and Language as Task Specifiers

An important special case of plug-and-play foundation models is to use text commands or visual inputs as task specifiers to learn more robust, general, and multi-task policies[Ahn et al. 2022; Huang et al. 2022a; Brohan et al. 2022; Liu et al. 2022a]. For instance, a text description of "close the cabinet door" or a goal image with the cabinet door closed can serve as policy input to augment the current robot state. There are a few motivations behind this approach. First, using language and a goal image to specify a task provides richer information about the intended task rather than merely providing a scalar reward. Second, pretrained language models (equipped with prompting methods such as chain-of-thought) can decompose high-level tasks into lower-level instructions that are easier to execute [Ahn et al. 2022; Huang et al. 2022a; Jiang et al. 2022; Team et al. 2021]. Furthermore, pretrained vision-language models can enable language-conditioned agents to generalize to new instructions, scenes, and objects in navigation and manipulation tasks [Lynch and Sermanet 2020; Hill et al. 2020; Hao et al. 2020; Majumdar et al. 2020; Nair et al. 2022; Jang et al. 2022a; Ahn et al. 2022; Huang et al. 2022a; Khandelwal et al. 2022; Shridhar et al. 2022; Guhur et al. 2022; Shah et al. 2022], which has been a key challenge in robotics prior to their introduction [Zhu et al. 2018].

Using vision and language task specifiers to prompt for desirable agent behaviors requires additional data such as text descriptions or goal images of a given task (see challenges in Section 6.1). Moreover, prompting for desirable outcomes from a large language model has significant potential but is also an open problem in itself [Liu et al. 2023b], whose complexity is exacerbated in decision making scenarios with external entities and world dynamics (see Section 6.4).

4.3 Learning Representations for Sequential Decision Making

Unlike vision-language foundation models that can learn from a broad data collection $\mathcal{D}$ but lack the notion of decision making, foundation model techniques and architectures (as opposed to the

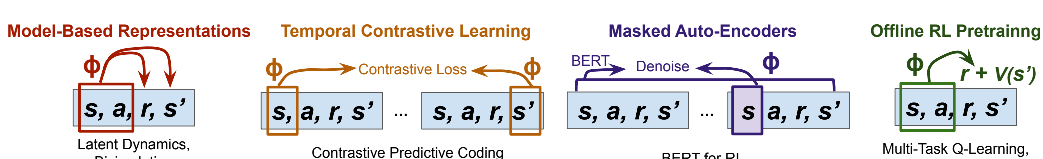

Fig. 4. Illustrations of different representation learning objectives such as model-based representations [Nachum and Yang 2021], temporal contrastive learning [Oord et al. 2018], masked autoencoders [Devlin et al. 2018], and offline RL [Kumar et al. 2022], on a trajectory $\tau \sim \mathcal{D}_{\mathrm{RL}}$ specifically devised for sequential decision making.

pretrained models themeselves) can be used to optimize objectives uniquely devised for sequential decision making on the basis of task-specific interactive data $\mathcal{D}_{\text{RL}}$. Figure 4 visually illustrates these representation learning objectives.

Model-based representations. Traditionally, representation learning for sequential decision making has been framed as learning a latent state or action space of an environment by "clustering" states and actions that yield similar transition dynamics [Dearden and Boutilier 1997; Andre and Russell 2002; Mannor et al. 2004; Abel et al. 2018; Gelada et al. 2019; Agarwal et al. 2021]. Similar to how foundation models can serve as generative models of world dynamics by maximizing $p(\tau)$ in Equation 18, foundation models can also serve as representation learners of world dynamics under the following objective:

$$p(\tau_{s,r}) = \Pi_{t=0}^{H} p(s_{t+1}, r_t | \tau_{<t}, s_t, a_t) = \Pi_{t=0}^{H} \mathcal{T}(s_{t+1} | \tau_{<t}, \phi(s_t), a_t) \cdot \mathcal{R}(r_t | \tau_{<t}, \phi(s_t), a_t). \tag{22}$$

Using this factorization for maximum likelihood estimation of $p(\tau_{s,r})$ using $\mathcal{D}_{\text{RL}}$ naturally leads to learning state representations $\phi(s)$ that "cluster" states with similar rewards and next state probabilities. One could also choose to maximize the likelihood of the next state *representations* as opposed to the next raw state, i.e., $\mathcal{T}(\phi(s_{t+1}) | \tau_{<t}, \phi(s_t), a_t)$ resulting in a latent dynamics model [Gelada et al. 2019]. Alternative learning objectives for $\phi(s)$ can be derived depending on how $\mathcal{T}(s_{t+1} | \tau_{<t}, \phi(s_t), a_t)$ is defined. For instance, $\mathcal{T}$ may be defined as an energy-based model:

$$\mathcal{T}(s_{t+1} | \tau_{<t}, \phi(s_t), a_t) \propto \exp\{\phi(s_{t+1})^\top f(\phi(s_t), a_t, \tau_{<t})\}, \tag{23}$$

where f is a trainable function that maps $\phi(s_t), a_t, \tau_{<t}$ to the same embedding space as ϕ . While Equation 22 learns state representations by modeling the forward dynamics, one can also learn state representations based on an *inverse* dynamics model [Pathak et al. 2017; Shelhamer et al. 2016] by predicting a_t from $\tau_{<t}, s_t, s_{t+1}$, thereby maximizing

$$p(\tau_a) = \Pi_{t=0}^{H} p(a_t | \tau_{<t}, \phi(s_t), \phi(s_{t+1})). \tag{24}$$

In addition to forward and inverse dynamics based representations, it is also possible to learn state representations derived from predicted value functions [Oh et al. 2017], curiosity metrics [Du et al. 2021], or other MDP-based similarity metrics such as bisimulation properties deduced from Bellman backups [Ferns et al. 2004; Castro and Precup 2010; Zhang et al. 2020]. The above representation learning objectives have mostly been considered under the Markovian setting, hence the dependence on $\tau_{<t}$ is often dropped. Though the Markovian assumption makes large sequence models seem less relevant, these representation learning objectives benefit from sequence modeling architectures in image-based domains that are generally non-Markovian.

Temporal contrastive learning. The model-based representation objectives above require strictly interleaved state-action-reward tuples in the training data $\mathcal{D}_{\text{RL}}$, which can preclude more flexible representation learning techniques that consider broader data sources, $\mathcal{D}$, such as YouTube videos (which can be thought of as state-only trajectories τ_s). Temporal contrastive learning such as CPC [Oord et al. 2018], on the other hand, can model more flexible sequence-level representations, and has been applied to playing games by watching YouTube videos [Aytar et al. 2018]. Specifically, in temporal contrastive learning, observations that are closer temporally (e.g., observations that belong to the same trajectory) are encouraged to have similar representations. Given a sub-trajectory $\tau_{t:t+h}$, one can learn $\phi(s)$ by minimizing a contrastive loss between $\phi(s_t)$ and $\phi(s_{t+i})$:

$$-\phi(s_{t+i})^\top W_i \phi(s_t) + \log \mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W_i \phi(s_t)\}]. \tag{25}$$

where $i = 1, \ldots, h$, W_i is a learnable weight matrix, and ρ is some non-trainable prior distribution. Note that the temporal contrastive learning in Equation 25 bears resemblance to learning an energy-based dynamics model in Equation 23, as established in prior work [Nachum and Yang 2021; Nguyen et al. 2021].

Masked autoencoders. When a trajectory $\tau = (s_0, a_0, r_0, ..., s_H, a_H, r_H)$ from $\mathcal{D}_{\mathrm{RL}}$ is treated as a flattened sequence, BERT-style denoising autoencoding objectives can be applied to the sequence to learn representations of states, actions, rewards, and dynamics through specific choices of masking patterns [Yang and Nachum 2021; Liu et al. 2022c; Carroll et al. 2022; Seo et al. 2022a]. These methods learn representations $\phi(s)$ by first randomly masking a subset of tokens in τ to obtain $\hat{\tau}$, then pass the masked sequence $\hat{\tau}$ to a transformer, and finally reconstruct the masked portions of the original input $\bar{\tau}$ from the transformer output $F(\hat{\tau})$. The training objective, for instance, can be characterized as maximizing

$$p(\bar{\tau}|\hat{\tau}) = \Pi_{t=0}^{H} m_t p(\tau_t|\hat{\tau}) = \Pi_{t=0}^{H} m_t \frac{\exp\{F(\hat{\tau})_t^T \phi(s_t)\}}{\sum_s \exp\{F(\hat{\tau})_t^T \phi(s)\}}, \tag{26}$$

where for each masked input state s_t, a contrastive loss between its representation $\phi(s_t)$ and the transformer output at its sequential position $F(\hat{\tau})_t$ is applied. Unlike model-based representation learning approaches that explicitly model state transition probabilities, masked autoencoders can learn representations from a broader dataset that potentially has missing actions and rewards, while still being able to incorporate dynamics-based information in the learned representations.

Offline RL pretraining. When the downstream decision making tasks are to be trained with RL objectives, it might seem natural to apply similar RL objectives during pretraining when acquiring value-based representations [Mazoure et al. 2022; Ball et al. 2023]. At a high level, value-based pretraining encompasses any offline RL algorithms that have been pretrained on logged experience from one or more tasks relevant to the downstream interactive task of interest. Value-based pretraining has exhibited scaling capability in multi-task settings where state action spaces are similar (e.g., all of Atari games [Kumar et al. 2022]).

4.3.1 *Post Representation Learning: BC and RL Finetuning.*

Unlike generative foundation models that can directly produce action or next state samples, as in Section 3, foundation models as representation learners are only directed to extract representations of states, actions, and dynamics; hence they require additional finetuning or model-based policy optimization to achieve strong decision making performance. On the theoretical side, various works have focused on developing representation learning objectives that ensure downstream BC or policy/value-based RL finetuning using the pretrained representations are provably efficient [Jin et al. 2020; Nachum and Yang 2021; Zhang et al. 2022b; Pacchiano et al. 2022; Ren et al. 2022]. These analyses are generally based on properties of linear MDPs. For instance, one such assumption states that the state-action value function $Q^\pi(s, a)$ can be represented as a linear combination of features $\phi(s, a)$ under the linear MDP factorization $\mathcal{T}(s'|s, a) = \langle \phi(s, a), \theta(s') \rangle$ and $\mathcal{R}(s, a) = \langle \phi(s, a), \theta_r \rangle$, which ensures that standard policy and value based RL training can take place in the more compact representation space $\phi(s, a)$ as opposed to the original state-action space. Beyond providing compact state action spaces for policy and value-based model-free RL methods, pretrained representations can also simplify model learning and policy rollouts of model-based policy optimization [Silver et al. 2014; Oh et al. 2017; Hafner et al. 2019] as described in Section 3.3.

While representation learning objectives specifically devised for sequential decision making have theoretical benefits, it is less clear how these objectives can effectively incorporate broader and multi-task data when the underlying dynamics differ from that of the target task of interest. The recurring challenge of bridging learning from broad data $\mathcal{D}$ and task-specific data $\mathcal{D}_{\mathrm{RL}}$ will be further discussed in Section 6.1.

5 LARGE LANGUAGE MODELS AS AGENTS AND ENVIRONMENTS

We have seen that foundation models can characterize different components of a decision making process ($\mathcal{M}$), such as agent behaviors (A), world dynamics ($\mathcal{T}$), task specifiers ($\mathcal{R}$), and state (S) and action representations. In this section, we further consider a special case where pretrained large language models can serve as agents or environments. Treating language models as agents, on one hand, enables learning from environment feedback produced by humans, tools, or the real world, and on the other hand enables new applications such as information retrieval and web navigation to be considered under a sequential decision making framework. Language models can also be thought of as computational environments that take text as input and produce text as output, effectively supporting interactions with external prompts.

5.1 Interacting with Humans

Dialogue as an MDP. A piece of dialogue can be viewed as in alternating nteraction between a dialogue agent π and a human environment $\mathcal{M} = \mathcal{E}$, where a conversation $\tau_{<t} = \{e_0, a_1, e_1, ..., a_t\}$ consists of sentences a_i and e_i produced by π and $\mathcal{E}$ respectively. On the t-th turn, a state $s_t \in S$ captures the conversation history $s_t = \{\tau_{<t}, e_t\}$, an action $a_t \in A$ is an agent's response given this context, a next state $s_{t+1} \in S$ concatenates s_t with a_t and e_{t+1}, and a reward $r_t = \mathcal{R}(s_t, a_t)$ is produced. An agent π aims to maximize $\mathbb{E}_{e_0 \sim \mu, \pi, \mathcal{T}}[\sum_{t=0}^{H} \gamma^t \mathcal{R}(s_t, a_t)]$.

Optimizing dialogue agents. The application of large language models to dialogue generation is a natural one, as both the broad pretraining data $\mathcal{D}$ and the task-specific dialogue data $\mathcal{D}_{\text{RL}}$ are of the same text modality, which allows for task-specific finetuning using the same self-supervised loss as pretraining [Adiwardana et al. 2020; Roller et al. 2021; Nakano et al. 2021; Thoppilan et al. 2022]. Such an approach has achieved impressive performance as assessed by humans, under metrics including safety, sensibleness, interestingness, truthfulness, and helpfulness [Thoppilan et al. 2022; Bai et al. 2022]. Although human feedback was initially used to evaluate dialogue systems [Jiang et al. 2021b], it was soon incorporated as a reward signal for optimizing dialogue agents under the *reinforcement learning with human feedback* (RLHF) framework [Ouyang et al. 2022; OpenAI 2022; Bai et al. 2022, *inter alia*]. In practice, RLHF involves several stages: first, a pretrained language model is finetuned on dialogue data to provide an initial policy π; second, output from this model is ranked by human raters, which is then used to train a preference (reward) model $\mathcal{R}$; finally, the language model is finetuned using policy gradient in Equation 4 to maximize the reward given by the preference model. Other RL objectives such as Q-learning (Equation 5) and actor-critic (Equation 6) have also been used to enable dialogue agent to perform specific tasks, such as booking flights and selling items on Craigslist [Jaques et al. 2017; Verma et al. 2022; Snell et al. 2022b; Jang et al. 2022b; Snell et al. 2022a].

Limitations of dialogue agents. While using human feedback is a natural way to turn broad data $\mathcal{D}$ into task-specific data $\mathcal{D}_{\text{RL}}$, solely relying on human feedback to finetune a language model agent has a number of limitations. For instance, language models have been criticized for failing to access up-to-date information [Komeili et al. 2021], hallucinating facts [Maynez et al. 2020; Ji et al. 2022], and struggling to perform complex reasoning and mathematical calculations [Patel et al. 2021]. Such failure modes are unsuprising if these desired properties were never a part of the feedback the language model received. While one approach to mitigate such failure modes is to collect human feedback on each of the desired properties, leveraging tools and external entities that can automatically provide feedback is likely to be a more scalable and reliable approach.

5.2 Interacting with Tools

Language model agents that generate API calls (to invoke external tools and receive responses as feedback to support subsequent interaction) can be formulated as a sequential decision making problem analogous to the dialogue formulation in the previous section. Several tools such as search engines [Komeili et al. 2021; Thoppilan et al. 2022; Lazaridou et al. 2022; Shuster et al. 2022; Yao et al. 2022], calculators [Cobbe et al. 2021; Thoppilan et al. 2022], translators [Thoppilan et al. 2022], MuJoCo simulators [Liu et al. 2022d], scratch pads [Nye et al. 2021], computer memory [Schuurmans 2023], and program interpreters [Gao et al. 2022] have been used to augment language models in a supervised finetuning or prompting setting, where response from tools are used as additional inputs to the language model.

Limitations of tool use agents. Unlike dialogue systems, where the agent and environment take turns, tool-using agents need to additionally decide when to call external tools, which tools to use, and how to use these tools (e.g., reformulating query if results are not helpful), all of which pose additional challenges. Consequently, the supervised finetuning of tool-use agents requires significant human supervision through API call annotations. While prompting-based tool-use requires fewer examples, the specific prompts typically need to be hand-crafted for each tool [Schick et al. 2023]. Moreover, language models are known to be sensitive to the prompt formats in both the zero and few-shot settings [Jiang et al. 2020; Schick and Schütze 2021]. As a result, the communication between language models and external tools typically needs to be cleaned-up by a rule-based parser, which further complicates the prompting setup. Recently, Parisi et al. [2022] and Schick et al. [2023] have made progress on self-supervised learning of tool use with language models, training the language model to only an external tool if this leads to an improved response over the outcome predicted by language model alone. Nevertheless, none of the existing work considers tool use in an interactive setting where an agent can *iterate* on its behavior according to tool feedback to improve its tool-use ability.

Tools as interactive environments. It is challenging to scale supervised finetuning and prompting to a large number of tools with different uses and tools that return large amounts of feedback (e.g., hundreds of search results). One sensible way of tackling this challenge is to treat tools like web browsers as interactive environments, from which experience can be sampled by executing search queries [Nakano et al. 2021; Gur et al. 2022], and optimizing such queries via RL techniques such as policy gradient. Treating tools as interactive environments enables methods that require massive and efficient online simulator access (e.g., Monte Carlo Tree Search for AlphaGo) to be applied to a broader set of real-world problems, such as web navigation and information retrieval. Additionally, situating language models in true knowledge obtained from the environment better grounds the model, avoiding the the Dichotomy of Control problem (e.g., sequence models generating next states without respecting environment transitions) [Yang et al. 2022b].

5.3 Language Models as Environments

Prompting as an MDP. Iterative prompting can be characterized as an MDP that captures the interaction between a prompt provider π and a language model environment $\mathcal{E}$, where a prompt history $\tau_{<t} = \{e_0, a_1, e_1, ..., a_t\}$ consists of prompts a_i and language model outputs e_i produced by π and $\mathcal{E}$ respectively. Here, e_0 is the initial context to the language model. In the t-th turn, a state $s_t \in S$ captures the prompting history and the t-th language model responses $s_t = \{\tau_{<t}, e_t\}$, an action $a_t \in A$ is given by the prompt provider, a next state $s_{t+1} \in S$ is produced by concatenating s_t with a_t and the next response of the language model e_{t+1}, and a reward $r_t = \mathcal{R}(s_t, a_t)$ is emitted. An agent π aims to maximize $\mathbb{E}_{e_0\sim\mu,\pi,\mathcal{T}}[\sum_{t=0}^{H} \gamma^t \mathcal{R}(s_t, a_t)]$. In language model reasoning, for instance,

$\mathcal{R}(s_t, a_t) = 1$ if the language model's output successfully reaches a goal answer s_t (i.e., correct reasoning), and $\mathcal{R}(s_t, a_t) = 0$ otherwise.

Under this formulation, various schemes for language model prompting can be characterized by high-level actions that map input strings to desired output strings using the language model. For instance, such high-level actions include `DECOMPOSE` [Press et al. 2022], `RANK` [Kumar and Talukdar 2021], `DENOISE` [Shi et al. 2023], and `PARAPHRASE` [Jiang et al. 2021a]. These high-level actions can also be recursively composed to achieve more sophisticated iterative prompting schemes [Zhou et al. 2022]. Other prompting schemes such as `SUMMARIZE`, `PRUNE`, `SEARCH` can be considered for handling challenges such as overcoming long context lengths. Given that language models with auxiliary memory have been shown to emulate universal Turing machines [Schuurmans 2023], language models could ultimately serve as "computers" that also operate on human language with prompting as a flexible new form of programming language.

6 OPEN PROBLEMS, CHALLENGES, AND OPPORTUNITIES

6.1 How to Leverage or Collect Datasets

One key challenge in applying foundation models to decision making lies in the dataset gap: the broad datasets from vision and language $\mathcal{D}$ and the task specific interactive datasets $\mathcal{D}_{\mathrm{RL}}$ can be of distinct modalities and structures. For instance, when $\mathcal{D}$ consists of videos, it generally does not contain explicit action labels indicating the cause-effect relationship between different frames, nor does it contain explicit reward labels indicating which videos are better than others, whereas actions and rewards are key components of $\mathcal{D}_{\mathrm{RL}}$. Despite this gap, broad video and text data can be made more task specific through post-processing ($\mathcal{D} \to \mathcal{D}_{\mathrm{RL}}$), leveraging hindsight relabeling of actions and rewards (e.g., using human feedback). Meanwhile, decision making datasets can be made more broad and general ($\mathcal{D}_{\mathrm{RL}} \to \mathcal{D}$) by combining a wide range of tasks-specific datasets (e.g., Gato). Below we provide a list of examples of $\mathcal{D}$ and $\mathcal{D}_{\mathrm{RL}}$ that can be used for research in foundation models for decision making, and propose additional approaches for bridging the gap between $\mathcal{D}$ and $\mathcal{D}_{\mathrm{RL}}$.

Existing vision and language datasets ($\mathcal{D}$). Vision and language datasets can be useful for decision making if they contain multiple modalities (e.g., aligned image and text pairs), (implicit) actions, movements, instructions, and notions of tasks. For instance:

- LAION-5B [Schuhmann et al. 2022] contains 5.85 billion CLIP-filtered text-image pairs.
- Egocentric 4D Perception (EGO4D) [Grauman et al. 2022] contains over 30k hours of time-aligned video in an inertial measurement unit (IMU) dataset of people's activities such as cooking, eating, and working at a computer in 4D (3D spatial and time).
- Something-Something V2 Dataset [Goyal et al. 2017] contains 220k short videos of people performing various tasks with everyday objects, such as putting on a hat and opening a bottle. These videos are annotated with action labels at the level of verb and noun phrases.
- HowTo100M [Miech et al. 2019] contains over 100 million video clips and descriptive captions, covering topics such as cooking, home improvement, and beauty.
- BigBench [Srivastava et al. 2022] is a dataset consisting of NLP tasks such as question answering, summarization, and conversation modeling. It also contains text-based games such as text navigation, Sudoku, and Taboo.

Existing decision making datasets ($\mathcal{D}_{\mathrm{RL}}$). Foundation models are currently relevant to decision making datasets that are larger-scale, multi-task, multi-modal, real-world based, and video or text based. For example:

- BabyAI [Chevalier-Boisvert et al. 2018] contains data in text-based games that require an agent to navigate in a 2D gridworld virtual environment and perform a variety of tasks.
- VirtualHome [Puig et al. 2018] contains over 15k simulated images and videos of indoor scenes, along with detailed information of the scenes and objects such as object shape, size, and material properties.
- RoboNet [Dasari et al. 2019] contains over 100k video clips of 7 robots over 100 camera viewpoints performing a variety of tasks in different environments.
- RL Unplugged [Gulcehre et al. 2020] is an offline RL dataset consisting of simulated locomotion, manipulation, and Atari games.
- Bridge Data [Ebert et al. 2021] contains 7,200 text-video demonstrations of a 6-dof WidowX250s robot arm performing 71 tasks across 10 kitchen-themed environments.
- MineDojo [Fan et al. 2022] contains 640k text-video pairs (16s in length), 7k Wiki pages, and 340k Reddit posts on Minecraft.
- RT-1 [Brohan et al. 2022] Robotics Transformer for Real-World Control at Scale (to be released).
- CACTI [Mandi et al. 2022]: A Framework for Scalable Multi-Task Multi-Scene Visual Imitation Learning (to be released).
- VIMA [Jiang et al. 2022] contains 650K successful trajectories of 17 simulated robotic manipulation tasks with interleaved language and image/video frames.

Bridging $\mathcal{D}$ and $\mathcal{D}_{\mathrm{RL}}$. To enable better datasets tailored for decision making, one can either increase the scale of $\mathcal{D}_{\mathrm{RL}}$ by large-scale logging and merging task-specific sets of interactive data, or by relabeling $\mathcal{D}$ with action and reward information. One could also consider augmenting $\mathcal{D}_{\mathrm{RL}}$ with meta data, such as informational and instructional texts and videos.

- Large-scale logging of interactions. Since many automatable tasks are currently conducted by humans (driving, navigating the web, writing code), it is possible to collect large amounts of data for sequential decision making by logging human behaviors. Similar to logged human conversations that are used to train dialogue agents, one can log "actions" such as keystrokes and mouse movements for training web navigating agents.
- Hindsight relabelling of existing data. Since many videos are already available on YouTube, it is possible to relabel the videos in hindsight with task descriptions and action information similar to Behbahani et al. [2019]; Shaw et al. [2022].
- Incorporating descriptions, instructions, and other task information. Since training a DQN Atari agent from scratch requires 7 GPU days, it is natural to consider whether information about an Atari game on the Internet (e.g., the Gameplay section of a game's Wikipedia page) could improve an agent's learning speed and sample efficiency.

6.2 How to Structure Environments and Tasks

Foundation models in vision and language can often solve a diverse set of tasks and generalize to new tasks in a few-shot or zero-shot manner [Radford et al. 2021; Alayrac et al. 2022; Brown et al. 2020; Chowdhery et al. 2022; Hoffmann et al. 2022]. Unlike vision and language where images or texts can serve as a universal task interface, decision making faces environment diversity where different environments operate under distinct state action spaces (e.g., the joint space and continuous controls in MuJoCo are fundamentally different from the image space and discrete actions in Atari), thereby preventing knowledge sharing and generalization. Below are some recent approaches to structuring environments and tasks so that foundation model architectures (e.g., Transformers) and large pretrained models (e.g., video diffusion) can be applied to decision making.

- **Universal encoding.** Similar to Reed et al. [2022] and Janner et al. [2021], all states, actions, and rewards across different environments and tasks can be encoded into universal tokens in a sequence modeling framework. However, such tokenization might not be able to preserve the rich knowledge and generalization abilities of pretrained vision and language models.

- **Text as environment.** Alternatively, one can convert environments with different state action spaces into text descriptions and use text as a universal interface to learn generalizable policies. For instance, when an observation is an image, one may use a caption model to convert the observation to text, or directly use ASCII characters to represent the observation as text. Text-as-environment and LM-as-policy have been evaluated on a variety of simple interactive games such as Spelling Bee, Sudoku, Chess, and Taboo [Srivastava et al. 2022], though there is still a substantial gap between large language models and state-of-the-art task-specific game-solving systems (e.g., AlphaGo) in these tasks. Text as environment also seems unnatural in visual perception based applications such as self-driving. Instead of using text as states and actions, one can also use text descriptions to specify tasks (rewards) [Ahn et al. 2022; Huang et al. 2022a; Brohan et al. 2022; Du et al. 2023b], avoiding the difficulties around reward shaping. Using text as a task specifier requires additional data to be collected, and still faces the challenge of incongruent state action spaces across tasks.
- **Video as policy and world model.** Finally, one can use image frames as a universal interface to represent state action spaces, and use videos to represent policies [Du et al. 2023b]. This allows policy learning to leverage web-scale pretrained text-to-video models. However, the mapping from videos to joint actions of individual agents still requires further training. This approach is further complicated by the computational difficulty of effective video generative modeling.

6.3 Improving Foundation Models

Long-context and External Memory Effective decision making often requires long context of the prior history of observations and actions. In contrast, existing approaches typically rely on transformers that have a bounded context length. To emulate general-purpose computations and decision making, properly incorporating interactions with external memory is important. One approach is to leverage prompting of intermediate computations [Schuurmans 2023; Giannou et al. 2023] to extend computational context, but this approach is difficult to implement in practice due to the sensitivity of language models on prompt selection and ways of parsing the output. Another interesting direction for future exploration is to incorporate retrieval of past observations to enable effective decision making [Borgeaud et al. 2021].

Combining multiple foundation models. Different foundation models capture different data modalities, such as visual, textual, and cross-modal representations of data. To effectively execute decision making across different environments, it is desirable to jointly leverage information across different models. One approach to compose models across different modalities is to graft them [Alayrac et al. 2022] on top of a single large language model. Alternatively, language can be used as a ubiquitous interface in which separate foundation models can communicate [Zeng et al. 2022]. Different foundation models can further communicate through iterative optimization [Li et al. 2022a]. A limitation of existing works is that they either require finetuning [Alayrac et al. 2022] or defined interfaces within which models can communicate [Zeng et al. 2022; Li et al. 2022a], which prevents novel combinations of foundation models from being easily composed at test-time in a free-form manner.

Grounding foundation models in the world. Foundation models are typically trained on Internet-scale data without knowledge of the physical world. To effectively execute actions produced by foundation models in the real world, it is important to ground these models in both the underlying geometry and physics of the world. One existing approach uses intermediate outputs from simulators as context for action generation [Liu et al. 2022d]. Alternatively, foundation model outputs could be scored and optimized using feedback from simulators [Li et al. 2022a]. Existing works assume access to a simulator of the operating environment, which is not available in the physical world. Constructing systems that more accurately ground predictions in the physical world is therefore an interesting area for future research.

6.4 Improving Decision Making

How to extract desirable behavior. One key aspect of foundation models for decision making lies in effectively adapting task-agnostic models into task-specific agents. Various approaches can be seen as ways to "control" foundation models to produce desirable behaviors for specific tasks. For instance, large-pretrained language models can be specialized to output desired sentences through instruction finetuning [Wei et al. 2021] or few-shot prompting [Brown et al. 2020]. For conditional generative modeling of behavior, language goals [Du et al. 2023b], image goals [Brohan et al. 2022], returns [Lee et al. 2022], environment constraints [Ajay et al. 2022], and expert demonstrations [Reed et al. 2022] have all been explored as s conditioning factor for finetuning or prompting schemes, so that the models can be "controlled".

Aside from goal or instruction conditioned finetuning or prompting, two types of "iterative" approaches have also been applied to elicit expert behavior. The first approach iterates through a set of chain-of-thought reasoning or computation steps [Nye et al. 2021; Wei et al. 2022b; Yang et al. 2022a], with the hope that a sequence model supervised to emit similar chain-of-thought steps will achieve better generalization. The second approach iterates through a set of improvement steps from less to more desirable behaviors, with the hope that a sequence model supervised on the improvement sequence can continue to regress on the improvement trend [Laskin et al. 2022; Liu et al. 2023a]. Both of these approaches, together with goal conditioned supervised learning, can help extract desirable behavior without requiring explicit finetuning with RL objectives.

Offline to online. While conditional generative modeling can elicit expert behavior as discussed above, directly finetuning foundation model agents using RL objectives such as policy gradient is another approach. One major challenge that has prevented wide real-world adoption of RL finetuning is the need for large online samples to ensure learning progress [Li 2019]. Nevertheless, in game settings where massive online access is available (e.g., Go, Chess, Shogi, Dota, Atari), RL methods have surpassed human performance. Instead of avoiding online access altogether through offline RL or conditional generative modeling, language models as interactive agents enables massive online access to environments that are highly scalable and available (e.g., search engines, databases, compilers). Developing infrastructures that enable software tools as environments, remote procedure calls as interactions, and foundation models as policies can have a large impact on a wide range of real-world applications.

7 DISCUSSION AND PERSPECTIVES

Foundation models have achieved remarkable success in emulating human intelligence at earlier stages of development: seeing, hearing, speaking, reading, and writing. To transform these basic human abilities to world-class expertise, humans spend tens of thousands of hours practicing through trial and error [Gladwell 2008], interacting with the external world, making mistakes, and learning from them. Foundation models for decision making offers a path to transform general artificial intelligence capabilities in vision, language, and world knowledge into next-level expert capabilities.

As well as achieving more sophisticated intelligence, foundation models can also characterize different components of a decision making system, such as generative models of behavior and the world (Section 3), representations of world knowledge (Section 4), and interactive agents or environments through the usage of language (Section 5). Despite the initial successes, foundation models for decision making inevitably faces significant challenges, such as the gap in data modalities, ambiguities around environment and task structures, and missing components in current foundation models and decision making paradigms (Section 6). We hope that this manuscript can serve as a stepping stone toward developing autonomous agents with next-level intelligence and more sophisticated capabilities.

ACKNOWLEDGMENTS

We thank Bo Dai and Douglas Eck for reviwing this manuscript.

REFERENCES

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. 2018. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*. PMLR, 10–19.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

Alekh Agarwal, Sham Kakade, and Lin F Yang. 2020a. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*. PMLR, 67–83.

Rishabh Agarwal, Marlos C Machado, Pablo Samuel Castro, and Marc G Bellemare. 2021. Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning. *arXiv preprint arXiv:2101.05265* (2021).

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. 2020b. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*. PMLR, 104–114.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. 2022. Beyond Tabula Rasa: Reincarnating Reinforcement Learning. *arXiv preprint arXiv:2206.01626* (2022).

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691* (2022). https://arxiv.org/abs/2204.01691

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. 2022. Is Conditional Generative Modeling all you need for Decision-Making? *arXiv preprint arXiv:2211.15657* (2022).

Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. 2020. Opal: Offline primitive discovery for accelerating offline reinforcement learning. *arXiv preprint arXiv:2010.13611* (2020).

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113* (2019).

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *NeurIPS* (2022). https://arxiv.org/abs/2204.14198

David Andre and Stuart J Russell. 2002. State abstraction for programmable reinforcement learning agents. In *Aaai/iaai*. 119–125.

Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando De Freitas. 2018. Playing hard exploration games by watching youtube. *Advances in neural information processing systems* 31 (2018).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862* (2022).

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795* (2022).

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023. Efficient Online Reinforcement Learning with Offline Data. *arXiv preprint arXiv:2302.02948* (2023).

Feryal Behbahani, Kyriacos Shiarlis, Xi Chen, Vitaly Kurin, Sudhanshu Kasewa, Ciprian Stirbu, Joao Gomes, Supratik Paul, Frans A Oliehoek, Joao Messias, et al. 2019. Learning from demonstration in the wild. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 775–781.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

Hans Georg Bock and Karl-Josef Plitt. 1984. A multiple shooting algorithm for direct solution of optimal control problems. *IFAC Proceedings Volumes* 17, 2 (1984), 1603–1608.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426* (2021).

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. 2022. When does return-conditioned supervised learning work for offline reinforcement learning? *arXiv preprint arXiv:2206.01079* (2022).

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. RT-1: Robotics Transformer for Real-World Control at Scale. *arXiv preprint arXiv:2212.06817* (2022).

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

Eduardo F Camacho and Carlos Bordons Alba. 2013. *Model predictive control*. Springer science & business media.

Micah Carroll, Orr Paradise, Jessy Lin, Raluca Georgescu, Mingfei Sun, David Bignell, Stephanie Milani, Katja Hofmann, Matthew Hausknecht, Anca Dragan, et al. 2022. Unimask: Unified inference in sequential decision problems. *arXiv preprint arXiv:2211.10869* (2022).

Pablo Castro and Doina Precup. 2010. Using bisimulation for policy transfer in MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. 2022b. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481* (2022).

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022a. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* (2022).

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2018. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272* (2018).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Jerry Plappert, Matthias and T̃worek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. 2019. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215* (2019).

Richard Dearden and Craig Boutilier. 1997. Abstraction and approximate decision-theoretic planning. *Artificial Intelligence* 89, 1-2 (1997), 219–283.

Marc Deisenroth and Carl E Rasmussen. 2011. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*. 465–472.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. 2002. Multiple model-based reinforcement learning. *Neural computation* 14, 6 (2002), 1347–1369.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence.

2023. PaLM-E: An Embodied Multimodal Language Model. In *arXiv preprint arXiv:2302.11111*.

Yilun Du, Chuang Gan, and Phillip Isola. 2021. Curious representation learning for embodied intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10408–10417.

Yilun Du, Shuang Li, and Igor Mordatch. 2020. Compositional Visual Generation with Energy Based Models. In *Advances in Neural Information Processing Systems*.

Yilun Du, Toru Lin, and Igor Mordatch. 2019. Model Based Planning with Energy Based Models. *CORL* (2019).

Yilun Du and Igor Mordatch. 2019. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689* (2019).

Yilun Du and Karthik Narasimhan. 2019. Task-Agnostic Dynamics Priors for Deep Reinforcement Learning. In *International Conference on Machine Learning*.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023a. Guiding Pretraining in Reinforcement Learning with Large Language Models. *arXiv preprint arXiv:2302.06692* (2023).

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 2023b. Learning Universal Policies via Text-Guided Video Generation. *arXiv e-prints* (2023), arXiv–2302.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. 2021. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396* (2021).

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *arXiv preprint arXiv:2206.08853* (2022).

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. 2018. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101* (2018).

Norm Ferns, Prakash Panangaden, and Doina Precup. 2004. Metrics for Finite Markov Decision Processes.. In *UAI*, Vol. 4. 162–169.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. 2022. Implicit behavioral cloning. In *Conference on Robot Learning*. PMLR, 158–168.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. PAL: Program-aided Language Models. https://doi.org/10.48550/ARXIV.2211.10435

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. 2019. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*. PMLR, 2170–2179.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. 2023. Looped Transformers as Programmable Computers. *arXiv preprint arXiv:2301.13196* (2023).

Malcolm Gladwell. 2008. *Outliers: The story of success.* Little, Brown.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. 5842–5850.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E Turner, and Sergey Levine. 2016. Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247* (2016).

Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. 2022. Instruction-driven history-aware policies for robotic manipulations. *arXiv preprint arXiv:2209.04899* (2022).

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio Gómez, Konrad Zolna, Rishabh Agarwal, Josh S Merel, Daniel J Mankowitz, Cosmin Paduraru, et al. 2020. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 7248–7259.

Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2022. Understanding HTML with Large Language Models. *arXiv preprint arXiv:2210.03945* (2022).

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* (2019).

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2020. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193* (2020).

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13137–13146.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. 2020. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382* (2020).

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556* (2022).

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207* (2022).

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022b. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022).

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

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2022b. GPT-Critic: Offline Reinforcement Learning for End-to-End Task-Oriented Dialogue Systems. In *International Conference on Learning Representations*. https://openreview.net/forum?id=qaxhBG1UUaS

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. 2022. Planning with Diffusion for Flexible Behavior Synthesis. *arXiv preprint arXiv:2205.09991* (2022).

Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems* 34 (2021), 1273–1286.

N. Jaques, S. Gu, D. Bahdanau, J. M. Hernandez-Lobato, R. E. Turner, and D. Eck. 2017. Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. *International Conference on Machine Learning (ICML)* (2017).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *Comput. Surveys* (2022).

Haoming Jiang, Bo Dai, Mengjiao Yang, Tuo Zhao, and Wei Wei. 2021b. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. *arXiv preprint arXiv:2102.10242* (2021).

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094* (2022).

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021a. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics* 9 (2021), 962–977.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438. https://doi.org/10.1162/tacl_a_00324

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR, 2137–2143.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. 2019. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374* (2019).

Sham M Kakade. 2001. A natural policy gradient. *Advances in neural information processing systems* 14 (2001).

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*. PMLR, 651–673.

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14829–14838.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational diffusion models. *Advances in neural information processing systems* 34 (2021), 21696–21707.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Levente Kocsis, Csaba Szepesvári, and Jan Willemson. 2006. Improved monte-carlo search. *Univ. Tartu, Estonia, Tech. Rep* 1 (2006).

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566* (2021).

Ilya Kostrikov, Denis Yarats, and Rob Fergus. 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649* (2020).

Aviral Kumar, Rishabh Agarwal, Xinyang Geng, George Tucker, and Sergey Levine. 2022. Offline Q-Learning on Diverse Multi-Task Data Both Scales And Generalizes. *arXiv preprint arXiv:2211.15144* (2022).

Aviral Kumar, Xue Bin Peng, and Sergey Levine. 2019. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465* (2019).

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. *arXiv preprint arXiv:2106.01751* (2021).

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 5639–5650.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. 2022. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215* (2022).

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115* (2022).

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, et al. 2022. Multi-Game Decision Transformers. *arXiv preprint arXiv:2205.15241* (2022).

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858* (2022).

Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. 2022a. Composing Ensembles of Pre-trained Models via Iterative Consensus. *arXiv preprint arXiv:2210.11522* (2022).

Shuang Li, Xavier Puig, Yilun Du, Clinton Wang, Ekin Akyurek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. 2022b. Pre-trained language models for interactive decision-making. *arXiv preprint arXiv:2202.01771* (2022).

Yuxi Li. 2019. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973* (2019).

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

Fangchen Liu, Hao Liu, Aditya Grover, and Pieter Abbeel. 2022c. Masked Autoencoding for Scalable and Generalizable Decision Making. *arXiv preprint arXiv:2211.12740* (2022).

Hao Liu, Lisa Lee, Kimin Lee, and Pieter Abbeel. 2022a. Instruction-Following Agents with Jointly Pre-Trained Vision-Language Models. *arXiv preprint arXiv:2210.13431* (2022).

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Languages are Rewards: Hindsight Finetuning using Human Feedback. *arXiv preprint arXiv:2302.02676* (2023).

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022b. Compositional Visual Generation with Composable Diffusion Models. *arXiv preprint arXiv:2206.01714* (2022).

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. 2022d. Mind's Eye: Grounded Language Model Reasoning through Simulation. *arXiv preprint arXiv:2210.05359* (2022).

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247* (2021).

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. 2020. Learning latent plans from play. In *Conference on robot learning*. PMLR, 1113–1132.

Corey Lynch and Pierre Sermanet. 2020. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648* (2020).

Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. 2022. Zero-Shot Reward Specification via Grounded Natural Language. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*.

Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*. Springer, 259–274.

Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. 2022. CACTI: A Framework for Scalable Multi-Task Multi-Scene Visual Imitation Learning. *arXiv preprint arXiv:2212.05711* (2022).

Shie Mannor, Ishai Menache, Amit Hoze, and Uri Klein. 2004. Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the twenty-first international conference on Machine learning*. 71.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

Bogdan Mazoure, Benjamin Eysenbach, Ofir Nachum, and Jonathan Tompson. 2022. Contrastive Value Learning: Implicit Models for Simple Offline RL. *arXiv preprint arXiv:2211.02100* (2022).

Vincent Micheli, Eloi Alonso, and François Fleuret. 2022. Transformers are sample efficient world models. *arXiv preprint arXiv:2209.00588* (2022).

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. 2017. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems* 30 (2017).

Ofir Nachum and Mengjiao Yang. 2021. Provable representation learning for imitation with contrastive fourier features. *Advances in Neural Information Processing Systems* 34 (2021), 30100–30112.

Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. 2018. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7559–7566.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. 2020. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359* (2020).

Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. 2022. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*. PMLR, 1303–1315.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. 2021. Temporal predictive coding for model-based planning in latent space. In *International Conference on Machine Learning*. PMLR, 8130–8139.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114* (2021).

Junhyuk Oh, Satinder Singh, and Honglak Lee. 2017. Value prediction network. *Advances in neural information processing systems* 30 (2017).

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

OpenAI. 2022. CHATGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

Sherjil Ozair, Yazhe Li, Ali Razavi, Ioannis Antonoglou, Aaron Van Den Oord, and Oriol Vinyals. 2021. Vector quantized models for planning. In *International Conference on Machine Learning*. PMLR, 8302–8313.

Aldo Pacchiano, Ofir Nachum, Nilseh Tripuraneni, and Peter Bartlett. 2022. Joint Representation Training in Sequential Tasks with Shared Structure. *arXiv preprint arXiv:2206.12441* (2022).

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255* (2022).

Keiran Paster, Sheila McIlraith, and Jimmy Ba. 2022. You Can't Count on Luck: Why Decision Transformers Fail in Stochastic Environments. *arXiv preprint arXiv:2205.15967* (2022).

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191* (2021).

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.

Jan Peters, Katharina Mulling, and Yasemin Altun. 2010. Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Dean A Pomerleau. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems* 1 (1988).

Dean A Pomerleau. 1989. *Alvinn: An autonomous land vehicle in a neural network*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY ....

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and Narrowing the Compositionality Gap in Language Models. *arXiv preprint arXiv:2210.03350* (2022).

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

Martin L Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. 2017. Imagination-augmented agents for deep reinforcement learning. *Advances in neural information processing systems* 30 (2017).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. 2022. Planning with Large Language Models via Corrective Re-prompting. *arXiv preprint arXiv:2211.09935* (2022).

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

Machel Reid, Yutaro Yamada, and Shixiang Shane Gu. 2022. Can Wikipedia Help Offline Reinforcement Learning? *arXiv preprint arXiv:2201.12122* (2022).

Tongzheng Ren, Chenjun Xiao, Tianjun Zhang, Na Li, Zhaoran Wang, Sujay Sanghavi, Dale Schuurmans, and Bo Dai. 2022. Latent Variable Representation for Reinforcement Learning. *arXiv preprint arXiv:2212.08765* (2022).

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761* (2023).

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 255–269. https://doi.org/10.18653/v1/2021.eacl-main.20

Juergen Schmidhuber. 2019. Reinforcement Learning Upside Down: Don't Predict Rewards–Just Map Them to Actions. *arXiv preprint arXiv:1912.02875* (2019).

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015a. Trust region policy optimization. In *International conference on machine learning*. PMLR, 1889–1897.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015b. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438* (2015).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Dale Schuurmans. 2023. Memory Augmented Large Language Models are Computationally Universal. *arXiv preprint arXiv:2301.04589* (2023).

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. 2022a. Masked world models for visual control. *arXiv preprint arXiv:2206.14244* (2022).

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. 2022b. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*. PMLR, 19561–19579.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. 2018. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1134–1141.

Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. 2022. Behavior Transformers: Cloning k modes with one stone. *arXiv preprint arXiv:2206.11251* (2022).

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. 2022. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *arXiv preprint arXiv:2207.04429* (2022).

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. 2022. VideoDex: Learning Dexterity from Internet Videos. *arXiv preprint arXiv:2212.04498* (2022).

Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. 2016. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307* (2016).

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. *arXiv preprint arXiv:2302.00093* (2023).

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*. PMLR, 894–906.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moỹa Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Ỹ-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. https://doi.org/10.48550/ARXIV.2208.03188

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *International conference on machine learning*. PMLR, 387–395.

Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. 2020. Parrot: Data-driven behavioral priors for reinforcement learning. *arXiv preprint arXiv:2011.10024* (2020).

Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. 2022a. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871* (2022).

Charlie Snell, Sherry Yang, Justin Fu, Yi Su, and Sergey Levine. 2022b. Context-aware language modeling for goal-oriented dialogue systems. *arXiv preprint arXiv:2204.10198* (2022).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligent analysis*, Vol. 2. Washington, DC., 2–6.

Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. 2022. PlaTe: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters* 7, 2 (2022), 4924–4930.

Richard S Sutton. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*. Elsevier, 216–224.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems* 34 (2021), 251–266.

Allison C Tam, Neil C Rabinowitz, Andrew K Lampinen, Nicholas A Roy, Stephanie CY Chan, DJ Strouse, Jane X Wang, Andrea Banino, and Felix Hill. 2022. Semantic exploration from language abstractions and pretrained representations. *arXiv preprint arXiv:2204.05080* (2022).

Tianxin Tao, Daniele Reda, and Michiel van de Panne. 2022. Evaluating Vision Transformer Methods for Deep Reinforcement Learning from Pixels. *arXiv preprint arXiv:2204.04905* (2022).

Yuval Tassa, Tom Erez, and Emanuel Todorov. 2012. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4906–4913.

DeepMind Interactive Agents Team, Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Felix Fischer, Petko Georgiev, Alex Goldin, Tim Harley, et al. 2021. Creating multimodal interactive agents with imitation and self-supervised learning. *arXiv preprint arXiv:2112.03763* (2021).

Gerald Tesauro. 1994. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation* 6, 2 (1994), 215–219.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 23–30.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

David Venuto, Sherry Yang, Pieter Abbeel, Doina Precup, Igor Mordatch, and Ofir Nachum. 2022. Multi-Environment Pretraining Enables Transfer to Action Limited Datasets. *arXiv preprint arXiv:2211.13337* (2022).

Siddharth Verma, Justin Fu, Mengjiao Yang, and Sergey Levine. 2022. Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning. *arXiv preprint arXiv:2204.08426* (2022).

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022).

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

Oskar Von Stryk. 1993. Numerical solution of optimal control problems by direct collocation. In *Optimal control*. Springer, 129–143.

Oskar Von Stryk and Roland Bulirsch. 1992. Direct and indirect methods for trajectory optimization. *Annals of operations research* 37, 1 (1992), 357–373.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. 2016. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763* (2016).

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents. *arXiv preprint arXiv:2302.01560* (2023).

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361* (2019).

Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. 2022. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173* (2022).

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. 2021. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems* 34 (2021), 27395–27407.

Mengjiao Yang and Ofir Nachum. 2021. Representation matters: offline pretraining for sequential decision making. In *International Conference on Machine Learning*. PMLR, 11784–11794.

Mengjiao Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. 2022a. Chain of thought imitation with procedure cloning. *arXiv preprint arXiv:2205.10816* (2022).

Mengjiao Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. 2022b. Dichotomy of control: Separating what you can control from what you cannot. *arXiv preprint arXiv:2210.13435* (2022).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. https://doi.org/10.48550/ARXIV.2210.03629

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022). https://arxiv.org/abs/2204.00598

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2020. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742* (2020).

Qihang Zhang, Zhenghao Peng, and Bolei Zhou. 2022a. Learning to drive by watching youtube videos: Action-conditioned contrastive policy pretraining. In *European Conference on Computer Vision*. Springer, 111–128.

Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. 2022b. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*. PMLR, 26447–26466.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).

Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. 2020. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6538–6547.

Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas, et al. 2018. Reinforcement and imitation learning for diverse visuomotor skills. *arXiv preprint arXiv:1802.09564* (2018).