A Hypothetical Framework of Embodied Generalist Agent with Foundation Model Assistance

Anonymous CVPR submission

Paper ID *****

Abstract

Recent significant advancements in computer vision (CV) and natural language processing (NLP) have showcased the vital importance of leveraging prior knowledge obtained from extensive data for a generalist agent. However, there are limited explorations in utilizing internet-scale data to train embodied generalist agents. In this work, we propose a hypothetical framework that integrates the prior knowledge from foundation models into each component of the actor-critic algorithms for the generalist agents.

1. Introduction

Recently, the fields of Natural Language Processing (NLP) [3,5,21,27] and Computer Vision (CV) [7,15,22,23] have witnessed significant progress, primarily attributable to the ability to consume extensive datasets in Deep Learning (DL). Specifically, GPT models [3, 21] are built upon a large pre-trained corpus consisting of billions of texts from the internet, while Segment Anything [15] employs massive amounts of hand-labeled segmentation data. These large-scale models have demonstrated superior capabilities, including strong precision and generalization, by leveraging prior knowledge from substantial data [3, 23, 26]. However, for embodied AI, there are few works to introduce such prior knowledge to learn a generalist agent. For human beings, prior knowledge is fundamental as it facilitates the swift acquisition of new skills. For example, a child who has never witnessed the act of opening a door may fail to turn the doorknob. But one who has can quickly make it because he/she has got the prior that the doorknob can be rotated. Such prior knowledge acquired through observations and experiences can enable humans to rapidly adapt to diverse daily tasks, including opening a door, pick-and-placing, and so on. So initiating with reasonable priors and attaining generalization is significant, especially in unexpected scenarios. Therefore, data plays a critical role in developing generalist embodied agents by its prior knowledge.

In light of this, we examine three principal data sources in embodied AI and the prior knowledge they offer.

The first is simulated data, which relies on simulators resembling real-world scenarios. With this approach, through interacting with the environments, the agent can gather prior knowledge of the physics and geometry of the environments, including force and depth feedback [12,16]. And the agent can take the successfully learned policies in simulation as the prior knowledge of solving different tasks. When the domain gap between simulation and real world is minimal, such prior knowledge can facilitate rapid adaptation to real-world scenarios. However, it is almost impossible to simulate the whole physical world. Consequently, it is unlikely to develop a generalist agent through the simulated environmental data.

The second data source is self-embodiment, which can provide exact experiences of interactions between the current embodiment and the environment. Such data allow the agent to learn more quickly in current scenarios compared with learning from other types of data [1, 2]. However, the efficiency of collecting self-embodied data is limited when the agent policy cannot be deployed in a large scale. And the successful applications in NLP and CV emphasize the importance of scalable data for model learning. Therefore, this type of data cannot serve as the primary source of prior knowledge due to the limited scalability.

The third source of data is the internet-scale data from other embodiments. It has greater scalability and is more accessible. The large-scale data from the internet can provide abundant prior knowledge to accomplish various tasks, including visual semantic information, decision logic, and physical changes in environmental interactions [15, 21]. Learning from these data for embodied agents is akin to humans observing the behaviors of others and imagining how to complete daily tasks. Therefore, it is significant to apply the prior knowledge from the internet-scale data to learn policies for embodiments, which is the target of this work.

We propose a novel framework for embodied generalist agents from internet-scale data with foundation model as-

sistance. Since the data are from different embodiments, the foundation models can provide noisy value, policy and task-success reward functions. And a systematic learning process is necessary as it is unlikely to succeed in one trial due to the noise in priors, which can be solved by RL. Notably, policy gradient algorithms, such as SAC [10] or PPO [25], are widely recognized as effective methods for solving RL problems. Therefore, we propose to assist the actor-critic learning with the foundation model priors to learn embodied generalist agents efficiently. Specifically, the value and reward priors from the foundation models can estimate the states so as to provide advantages in actor-critic learning. And the policy prior can provide general strategies for solving tasks to guide the agent, which prevents amounts of random explorations. In this paper, we demonstrate how to obtain and utilize the priors from foundation models for actor-critic learning for embodied generalist agents.

2. Method

2.1. Problem Formulation

This work facilitates online policy learning with robotics embodiment across various scenarios. In reinforcement learning (RL), tasks are usually modeled as Markov decision processes (MDPs). To handle different tasks with one agent, Goal-conditioned RL (GCRL) settings have emerged as a more promising alternative compared to vanilla MDPs, which augment an additional goal for tasks. In this paper, we use language as the goal for the sake of its strong representational and generalization abilities. For better descriptions, we define the Goal-conditioned MDP (GCMDP) as a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_{|\mathcal{T}}, \mathcal{T})$. $\mathcal{S} \in R^m$ denotes the current state, $\mathcal{A}$ denotes the action space of the agent, and $\mathcal{P} = \Pr\{s_{t+1}|s_t, a_t\}$ denotes the transition probabilities. $\mathcal{T}$ is the task identifier, which is instantiated by language in this work. $\mathcal{R}_{|\mathcal{T}}$ denotes the rewards conditioned by tasks.

2.2. Actor-Critic Learning by Foundation Priors

The actor-critic algorithm, a temporal difference (TD) version of policy gradient algorithms, comprises a policy model π and a value model $\mathcal{V}$. Our basic idea is to provide appropriate priors to each module of the actor-critic algorithm through foundation model assistance.

Regarding the critic, we propose to approximate values and rewards based on a value foundation model $M_{\mathcal{V}}(s|\mathcal{T}) : \mathcal{S} \times \mathcal{T} \rightarrow R$, and a reward foundation model $M_{\mathcal{R}}(s|\mathcal{T}) : \mathcal{S} \times \mathcal{T} \rightarrow \{0, 1\}$. Here the model $M_{\mathcal{V}}(s|\mathcal{T})$ estimates the values of states, and dense rewards of each transition can be derived from it. However, due to the inherent noise in the value prior, an additional signal is required to improve the accuracy of the estimates. To address this, we introduce the reward model $M_{\mathcal{R}}(s|\mathcal{T})$, which acts as a success discriminator, providing a task completion signal for the final state. Such reward models offer greater precision and can be easily acquired. In this way, we are able to infer the advantages of transitions based on the value and reward priors.

Regarding the actor, our approach focuses on enhancing policy learning by leveraging the advantages derived from the value and reward priors, as well as incorporating the guidance provided by policy priors. Without policy priors, the agent would need to rely on extensive trial and error during policy gradient learning to perform well from random behaviors. However, humans possess a general strategy or predefined routes for handling tasks, which can be considered as a form of policy prior. To introduce similar policy priors for guidance, we suggest employing a language-vision foundation model $M_{\hat{\tau}}(s_0|\mathcal{T}) : \mathcal{S} \times \mathcal{T} \rightarrow \hat{\tau}$, which generates latent videos of task-solving trajectories $\hat{\tau}$. Since the current embodiment data (eg, actions) cannot be accessed in the internet-scale data, we propose to train an inverse dynamics model $\rho(s_t, s_{s+1}) : \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{A}$ in Eq. (1) by the rolled-out trajectory τ in environments.

$$L^{\rho} = KL(\rho(s_t, s_{t+1}), a_t), s_t, s_{t+1} \in \tau. \tag{1}$$

Then the inverse model ρ can infer actions from the generated task-solving trajectory, surpassing the performance of random actions. Afterward, the action provided by the inverse model ρ can serve as a policy prior to guiding the policy π by KL divergence. As a result, we train the policy π through actor-critic learning, incorporating the KL alignment with the policy prior as depicted in Equation (2).

$$\begin{aligned} L^{\pi} &= KL(\pi(s_t|\mathcal{T}), \rho(s_t, \hat{s}_{t+1})) - \alpha A^{\pi} \log \pi(s_t|\mathcal{T}), \\ &\text{where } A^{\pi} = \beta M_{\mathcal{R}}(s_{t+1}) + (M_{\mathcal{V}}(s_{t+1}) - M_{\mathcal{V}}(s_t)), \\ &\text{and } s_t, s_{t+1} \in \tau, \hat{s}_{t+1} \in \hat{\tau}. \end{aligned} \tag{2}$$

, where $\alpha = 1, \beta = 100$ are trade-offs. To optimize the inverse model ρ and policy π, we minimize the training objectives L^{ρ} and L^{π} respectively.

The detailed training pipeline is in Alg.1. For a given task $\mathcal{T}$, a task-solving trajectory $\hat{\tau}$ is generated, and the agent takes the policy π to roll out a trajectory τ in the environments. Then we simultaneously train the inverse dynamics model ρ and the policy model π. Finally, we can obtain the trained policy model π for the current embodiment.

Algorithm 1 Policy Learning Guided by Foundation Priors

Given online replay buffer D, task horizon H, inverse dynamics model ρ, policy model π, Foundation models $M_{\mathcal{R}}, M_{\mathcal{V}}, M_{\hat{\tau}}$.
while not finished task $\mathcal{T}$ **do**
  Generate task-solving trajectory $\hat{\tau} = M_{\hat{\tau}}(s_0|\mathcal{T})$.
  Roll out trajectory τ in the environment with π, $D \leftarrow D \bigcup (\tau, \hat{\tau})$
  Train ρ and π by objective (1) and (2) respectively.
end while=0

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1, 5

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1, 5

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[4] Yilun Dai, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023. 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[6] Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards a unified agent with foundation models. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*. 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 5

[9] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *arXiv preprint arXiv:2206.08853*, 2022. 5

[10] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 2

[11] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 5

[12] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019. 1

[13] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Ijcai*, pages 4246–4247, 2016. 5

[14] Anssi Kanervisto, Stephanie Milani, Karolis Ramanauskas, Nicholay Topin, Zichuan Lin, Junyou Li, Jianing Shi, Deheng Ye, Qiang Fu, Wei Yang, et al. Minerl diamond 2021 competition: Overview, results, and lessons learned. *NeurIPS 2021 Competitions and Demonstrations Track*, pages 13–28, 2022. 5

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[16] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021. 1

[17] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022. 5

[18] Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-shot reward specification via grounded natural language. In *International Conference on Machine Learning*, pages 14743–14752. PMLR, 2022. 5

[19] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022. 5

[20] Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *International Conference on Machine Learning*, pages 3878–3887. PMLR, 2018. 5

[21] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 1

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[24] 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. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 5

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 5

[26] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1), 2019. 1

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[28] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023. 5

A. Related Work

Robotics Learning with Large-scale Data The ability to leverage generalized knowledge from large and varied datasets has been shown in the fields of CV and NLP, but this is absent in robotics currently. Some works [1, 6, 8, 11] utilize the large language model for embodied planning or semantic guidance for alignment. Apart from the large language model for planning, SayCan [1] also trains a value network to provide grounding connections to the physics environments. For low-level control to reach the goals, they collect large amounts of in-domain data for behavior cloning. However, we are not aimed at solving the task-planning problems in robotics. Instead, we focus on the low-level atomic tasks for each subgoal, which is challenging due to the various control embodiments and complex environments. And this kind of scope is omitted when researchers apply the foundation models to the robotics field for better generalization ability. Our framework is key to the low-level control issues in robotics with large amounts of cheap and diverse data. Moreover, we introduce the reward and value foundation models for reinforcement learning. Meanwhile, the demonstrations generated by the video foundation model give heuristic guidance for the learned policy.

Robotics Transformer [2] is built on a novel transformer architecture with large amounts of multi-modal data, which takes an image and language description as input and generates the arm actions for low-level tasks. Gato [24] tokenizes multi-modal, multi-task as well as multi-embodiment inputs and scales up a large transformer sequence model to learn a generalist agent through imitation learning. However, those methods require large amounts of in-domain data for behavior cloning.

Policy Learning with Video Generation Models Apart from the works based on imitation learning, some researchers make progress in leading text-guided video generation for learning a universal policy. ROSIE [28] performs aggressive data augmentation for existing robotic manipulation datasets through text-to-image diffusion models. But the video generation model is considered an augmentation tool for policy learning. Instead, UniPi [4] learns video generation and inverse dynamics models on large in-domain expert videos. The video model gives a demonstration of the current state, and then the agent takes actions generated by the inverse dynamics. However, such a learning paradigm is not scalable and transferable across different embodiments and environments for the sake of training the inverse dynamics model offline. Instead, we attempt to learn a policy model rapidly for the current embodiment after some trials with the help of foundation models.

Foundation Models for Reward/Value Predictions For reinforcement learning, reward and value terms can be also approximated through foundation models under large amounts of task-agnostic data. MineDoJo [9] learns a reward function on top of large pre-trained video-language models for Minecraft [13, 14]. But they train the policy with PPO [20, 25] and self-imitation [20]. Instead, we introduce the video generation model for trajectory guidance, which accelerates policy learning on a large scale. Some works [18, 19] provide language-conditioned reward functions to generate task reward signals based on offline datasets. In terms of value prediction, VIP [17] is the first to train a goal-conditioned value function on large-scale unlabeled videos. Most of the current reward/value foundation models are conditioned on the language goal, which is part of the GMDP and can be unified in our framework.