Towards a Continuously Improving, Ever Curious, Generally Capable Foundational Embodied Agent Model

RUIJIE ZHENG, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF MARYLAND

1 Introduction

With the rapid development of foundational models, we now have dramatically expanded our capabilities to understand multimodal sensory inputs such as language, images, videos, and audios within open-world environments. This advancement sets the stage for the next goal in AI: **the development of interactive embodied foundational agent models**. These autonomous agents aim to not only perceive but also act within their environment and solve more complex sequential decision making tasks. This includes but is not limited to autonomous agent for tool manipulation (for example, an agent that could automatically file taxes), robotics (including but not limited to humanoid robots, robotic dogs, mobile manipulation and aerial robot), autonomous driving, and gaming agents.

Moving forward, I envision that a data-driven learning approach similar to the next token prediction for the pretraining language model could close the loop of perception and action, enabling the development of embodied AI foundational agents. However, unlike language models, the development of embodied foundational agents faces many unique challenges.

- Embodied agents—from gaming agents to autonomous vehicles and humanoid robots—operate within varied and complex action spaces. It is crucial to develop foundational embodied agent models that can effectively utilize data from these diverse embodiments.
- Unlike the language model, where we could easily get a huge amount of data by scraping the Internet for pretraining, data is much more expensive to collect for embodied agents, and the pretraining tasks are also not quite diverse enough to scale up the data.
- Since these agents will be applied in various scenarios, it is hard to enumerate all settings and train models on every single task, so it is essential for those models to have the ability to self improve and adapt to new tasks or situations by bootsraping from past sub-optimal trajectories.

To develop embodied foundational agents, we need to develop a principled approach to scale up task and data generation, as well as to develop pretraining algorithms that can absorb large amounts of data from diverse tasks and embodiments, adapting easily to unseen tasks with few-shot demonstrations or online interactions. Below, I will outline my current research and provide a roadmap for my future research agenda on foundational embodied agent models.

2 Overview of Current Research

In my research, I have addressed the challenge of using deep reinforcement learning (RL) to solve tasks based on raw pixel observations, such as images. Deep RL enables the embodied agent to effectively learn from historical suboptimal trajectories to achieve a better policy. However, its adoption in real-world scenarios has been limited due to its sample inefficiency, thanks to the entanglement of representation learning with credit assignment and exploration problems in sequential decision-making. My work has specifically focused on enhancing the sample efficiency of visual RL algorithms from two distinct angles.

I. Representation learning In our NeurIPS 2023 paper, TACO [17], we introduce a novel temporal contrastive learning approach for sequential decision-making tasks. Instead of directly modeling the transition dynamics, which can lead to representational collapse, TACO focuses on optimizing the mutual information between representations of current states paired with action sequences $[z_t, u_t, ..., u_{t+K-1}]$ and representations

of the corresponding future states z_{t+K} through a contrastive learning objective. TACO acts as a representation learning module compatible with any visual RL algorithm. We empirically show that it could significantly enhance sample efficiency and performance for both online and offline visual RL algorithms across the Deepmind Control Suite [9] and MetaWorld [14], two widely studied continuous control benchmarks.

II. Exploration-exploitation trade-off via dormant ratio Another of my recent works, DrM [13], provides a novel insight on the relationship between the agent's exploration behaviors with dormant ratio, an intrinsic measure of the agent policy network's "activity level". A low dormant ratio correlates with the agent's physical ability to actively explore the environment, whereas a high dormant ratio correlates with the agent's immobility. Based on this insight, we propose DrM, which aims to actively reduce the agent's dormant ratio and use it as a signal to guide the exploration-exploitation tradeoff. DrM significantly enhaces SoTA visual RL algorithms across a variety of challenging tasks, making us one step closer to end-to-end real-world visual RL applications.

Beyond the foundational research on how to enhance the sample efficiency of deep visual RL algorithms, another important question that I have been studied is how we could leverage large offline pretraining dataset so that for a new downstream task, the agent could adapt with only a few expert demonstrations or few online interaction steps using RL. Towards this objective, my previous research has considered the following two pretraining objectives.

1) Pretraining generalizable state representation Advancing upon TACO, we propose Premier-TACO, a multitask feature representation learning approach designed to improve few-shot policy learning efficiency. Compared with TACO, Premier-TACO incorporates a novel negative example sampling strategy tailored towards multitask pretraining. This strategy is crucial in significantly boosting TACO's computational efficiency and performance, making large-scale multitask offline pretraining feasible. With empirical evaluation under a diverse set of continuous control benchmarks including Deepmind Control Suite, MetaWorld, and LIBERO, we demonstrate Premier-TACO's effectiveness in pretraining visual representations, significantly enhancing few-shot imitation learning performance under unseen tasks.

2) Pretraining temporally extended action abstractions Another of my recent work, PRISE [15], approaches multitask pretraining from a different angle. Given a pretraining dataset of demonstrations from multiple tasks over a continuous action space, PRISE studies the problem of learning temporally extended action primitives, i.e., skills, to improve downstream few-shot imitation learning by capitalizing on a novel connection to NLP pretraining methodology. We demonstrate that by embedding continuous actions into discrete codes and applying a popular NLP tokenization method, Byte Pair Encoding (BPE)—commonly used in LLM pretraining—to these codes, we can identify variable-timespan action primitives that enable efficient downstream imitation learning.

In addition to the topics above, I have also done research related to model-based RL [18, 11] and transfer learning in RL [12, 8], leveraging world-models for efficient policy learning, and robust RL [7, 4, 6, 5] to develop robust policies under perturbation.

3 Future Research Agenda

Given my research in deep reinforcement learning and self-supervised pretraining for sequential decision-making, I plan to advance my research on embodied foundational agents with the following objectives: enabling the next-generation embodied agent model to be pretrained on a broader range of procedurally generated tasks and data, incorporating diverse embodiments, equipping the agent with the capability to self-improve efficiently through trial and error, and also reducing its inference costs.

I. Aligning Vision Language Model (VLM) with robotics data through temporal action tokenization Aligning multimodal language models with robotics data for low-level control has been highly successful,

as exemplified by Google's recent development of the RT-2 vision-language-action model [2], which finetunes a vision-language model using a large robotics dataset. However, one key bottleneck of the current approach is its slow inference speed, as robots need to execute actions at a high frequency by querying large models. Additionally, scaling up the data to accommodate diverse embodiments with heterogeneous action spaces presents a significant challenge.

To address the issues above, my temporal action tokenization work, `PRISE` [15], could potentially play a crucial role here. Instead of finetuning with raw action spaces, we could finetune VLMs with `PRISE` pretrained temporally extended action tokens, where each token corresponds to a **sequence** of closed-loop policies. This enables the robot to significantly improve inference speed without distilling the model, as it no longer needs to query the large model at each timestep. Furthermore, it would also be interesting for my future work to advance the `PRISE` tokenization mechanism to accommodate diverse heterogeneous action spaces, allowing us to construct unified action tokens across different embodiments and enhance the scalability of the RT-2 approach for incorporating data from various physical embodiments.

II. Self-improving agent through offline-to-online adaptation / in-context learning The current large embodied foundational models are primarily limited to imitation learning using large human-teleoperated demonstration datasets. These agents intrinsically lack the ability to self-improve in sub-optimal scenarios, such as when faced with unfamiliar tasks or significant changes in visual observations. In such cases, bootstrapping from their past non-successful trajectories through trial and error is essential, posing a unique challenge compared to existing LLM/VLMs.

Toward this goal, I am planning to conduct my research in two directions. First, I will explore how we can efficiently do online policy adaptation by leveraging knowledge pretrained on large offline datasets. `Premier-TACO` [16] demonstrated that pretraining a generalizable state representation can significantly enhance downstream performance. Going beyond state representation, we could also pretrain a world model, which could then be used for model-based policy optimization during downstream adaptation time. An important research question to address here is how to learn an accurate, universal world model and whether this could be linked to the recent development of large video prediction/representation models [3, 1], leveraging the extensive knowledge of these models learned from internet-scale data.

On the other hand, while RL can still be sample inefficient, the trial and error phase might alternatively be approached through in-context learning. During training, a large transformer model is given a sequence of trajectories with ascending rewards, learning to predict actions from better trajectories based on poorer ones. During evaluation, historical trajectories can simply be input into the context window, allowing the transformer to strategize improvements for subsequent trajectories. By curating a large and diverse dataset from different tasks/agents and training/finetuning a large transformer model on it, we could potentially unlock the agent's self-improvement capability without relying on traditional RL.

III. Pipeline of scaling up data for embodied agents via LLM A large capacity foundational model cannot be learned without large, diverse data. Recent works [10] have utilized the coding capabilities of LLMs to create new tasks in simulation to train a generalizable multitask policy. However, instead of generating all tasks at once, task generation should ideally be procedurally generated, tailored to the agent's capabilities. Thus, developing an iterative process would be beneficial, where at each round, the LLM is informed of the agent's current capabilities by analyzing its historical success rates on previous pool of tasks. The LLMs would then generate new tasks to enhance the agent's abilities, with the agent learning from these newly coded tasks. This process would continue, allowing the LLMs to progressively come up with new tasks to continuously expand the agent's capabilities. I envision this could potentially be a generic pipeline for scaling up data across various sequential decision-making applications, such as robotic manipulation, humanoid control, gaming agents, and autonomous tools manipulation agents.

[1] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video, 2024.

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[3] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. de Freitas, S. Singh, and T. Rocktäschel. Genie: Generative interactive environments, 2024.

[4] Y. Liang, Y. Sun, R. Zheng, and F. Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22547–22561. Curran Associates, Inc., 2022.

[5] Y. Liang, Y. Sun, R. Zheng, X. Liu, B. Eysenbach, T. Sandholm, F. Huang, and S. M. McAleer. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Y. Sun, R. Zheng, P. Hassanzadeh, Y. Liang, S. Feizi, S. Ganesh, and F. Huang. Certifiably robust policy learning against adversarial multi-agent communication. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Y. Sun, R. Zheng, Y. Liang, and F. Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In *International Conference on Learning Representations*, 2022.

[8] Y. Sun, R. Zheng, X. Wang, A. E. Cohen, and F. Huang. Transfer RL across observation feature spaces via model-based regularization. In *International Conference on Learning Representations*, 2022.

[9] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller. Deepmind control suite, 2018.

[10] L. Wang, Y. Ling, Z. Yuan, M. Shridhar, C. Bao, Y. Qin, B. Wang, H. Xu, and X. Wang. Gensim: Generating robotic simulation tasks via large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] X. Wang, R. Zheng, Y. Sun, R. Jia, W. Wongkamjan, H. Xu, and F. Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Y. Wei, Y. Sun, R. Zheng, S. Vemprala, R. Bonatti, S. Chen, R. Madaan, Z. Ba, A. Kapoor, and S. Ma. Is imitation all you need? generalized decision-making with dual-phase training. *arXiv preprint arXiv:2307.07909*, 2023.

[13] G. Xu, R. Zheng, Y. Liang, X. Wang, T. J. Zhecheng Yuan, Y. Luo, X. Liu, J. Yuan, P. Hua, S. Li, Y. Ze, H. D. III, F. Huang, and H. Xu. Drm: Mastering visual reinforcement learning through dormant ratio minimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.

[15] R. Zheng, C.-A. Cheng, H. D. III, F. Huang, and A. Kolobov. Prise: Learning temporal action abstractions as a sequence compression problem, 2024.

[16] R. Zheng, Y. Liang, X. Wang, S. Ma, H. D. III, H. Xu, J. Langford, P. Palanisamy, K. S. Basu, and F. Huang. Premier-taco is a few-shot policy learner: Pretraining multitask representation via temporal action-driven contrastive loss, 2024.

[17] R. Zheng, X. Wang, Y. Sun, S. Ma, J. Zhao, H. Xu, H. D. III, and F. Huang. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] R. Zheng, X. Wang, H. Xu, and F. Huang. Is model ensemble necessary? model-based RL via a single model with lipschitz regularized value function. In *The Eleventh International Conference on Learning Representations*, 2023.