Foundation Reward Models for General Robot Skill Acquisition

Yecheng Jason Ma, University of Pennsylvania

The long-term goal of my research is to develop generalist robots that can learn and perform a wide variety of tasks in unstructured environments. This level of robot versatility and generality could provide huge value through applications such as home assistants for the elderly or better automation for manufacturing. The sheer diversity of real-world robotic tasks, however, makes it impractical to hand-design or learn a new controller for every robot task. Instead, we must seek to develop robotic *foundation models* capable of solving a variety of tasks without explicit task-specific supervision [1, 3, 5]. Foundation models, such as GPT-4 [26], have shown that models trained on vast quantities of data can exhibit generalist capabilities. However, these existing foundation models are largely trained using internet-scale datasets, a luxury unlikely to be parallelled in the near future for robotics. This unique challenge means that we must rethink the recipe for training foundation models that can enable general robot autonomy, prompting the central question of my research: ***what are the key ingredients of robotic foundation models and how should we train them?***

I posit that foundation models for embodied intelligence must be capable of learning from **offline non-robot data** and provide **actionable information** for robots to learn and adapt skills in new environments. These desiderata can address the lack of internet-scale robot data as the robot can acquire useful knowledge elsewhere that will guide them to autonomously solve new tasks. My research to-date focuses on learning and deploying actionable foundation models for robotics, centered around the theme of learning ***foundation reward models*** for robot skill acquisition; (shaped) reward functions simultaneously specify the task objective (the what) and provide detailed feedback (the how) for skill learning. Effective large-scale reward learning algorithms can therefore bootstrap skill acquisition across tasks, embodiments, and observation inputs. I have worked on three complementary thrusts of foundation reward learning by leveraging data sources and models designed primarily for non-robot applications: (1) designing reward functions with large language models [21], (2) pre-training universal value functions from human videos [20], and (3) multi-modal value pre-training from multi-modal video data [19]. Collectively, my existing research presents a versatile toolkit of large-scale, foundational reward learning algorithms: they can learn entirely from non-robot data, handle tasks specified in image and language, express reward as black-box neural network as well as interpretable code, and support learning in the real world and simulation.

Reward Design via Large Language Models. Sim-to-real reinforcement learning is a promising paradigm for learning complex robot skills [23, 27, 33]. However, designing an effectively reward function for the simulation environment is notoriously difficult, time consuming, and prone to error [4, 14]. Given that reward design is a non-differentiable optimization problem in the space of programs, prior approaches have largely resorted to zeroth-order black-box optimization method such as evolutionary search [9, 25] or prompting large language models [37]. However, these prior approaches require extensive task-specific algorithmic design, struggling to scale to more difficult tasks. Leveraging the remarkable code generation capability of state-of-the-art large language models [26], I developed `Eureka` [21], the first fully-automated reward design algorithm that can generate reward functions comparable to human designed ones. The key idea of `Eureka` is to implement an evolutionary algorithm within the context of a LLM, progressively discovering more effective reward functions. Importantly, `Eureka` is free of any robot- or task-specific prompt or template engineering and can generate effective reward functions across a diverse set of 10 robot embodiments and 30 distinct tasks (Figure 1 Right). This streamlined design primes `Eureka` as a versatile reward design tool for new robot task in simulation, jumpstarting the sim-to-real design process. Recently, we have extended the `Eureka` framework to perform environment design [16] in `Eurekaverse` and to automate the whole design process of sim-to-real transfer [22] in `DrEureka`, enabling learning of novel skills such as a quadruped robot balancing on a yoga ball.

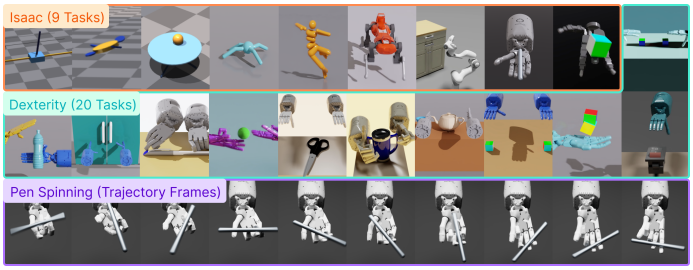

Fig. 1: `Eureka` can generate reward functions comparable to human-designed ones on a majority of tasks in a diverse suite of 10 robots and 30 tasks, including the novel dexterous pen spinning task.

Universal Value Pre-Training from Human Videos. Videos of humans accomplishing daily tasks are abundant on the internet. But how can these freely available, "in-the-wild" videos help robotics? Observing that human videos inherently exhibit goal-directed behavior, in my research, I proposed using them as transition data for offline goal-conditioned reinforcement

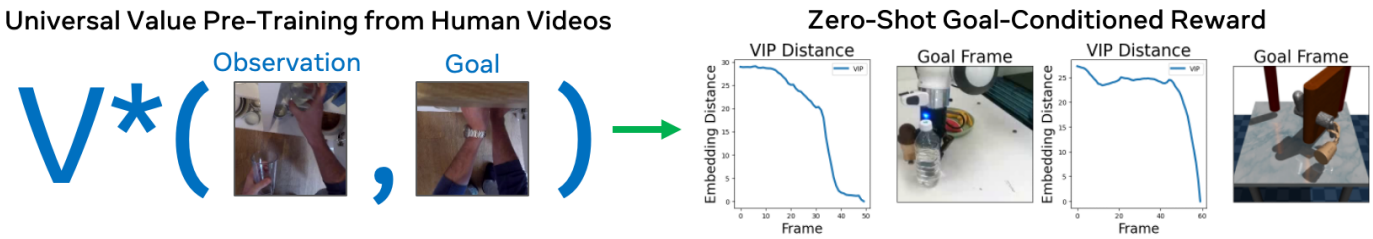

Fig. 2: VIP can zero-shot generate dense and smooth rewards for unseen robot tasks.

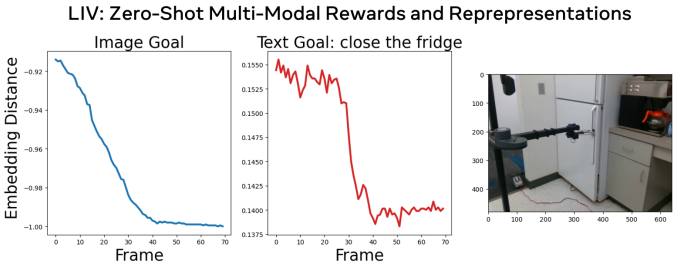

Fig. 3: LIV can zero-shot predict values with respect to either image or language goal.

learning (GCRL) to learn *universal value functions* [29] that can predict values and thereby rewards for any task. Drawing from my earlier research on dual optimization methods for offline reinforcement learning [17, 18], I delopved value-implicit pre-training (VIP [20]), which learns a visual representation wherein the embedding distance captures the optimal goal-conditioned value function. Pre-trained on a large number of in-the-wild egocentric human videos from Ego4d [12], VIP can produce zero-shot dense reward signals for a variety of unseen robot tasks specified via image goals (Figure 2). In both simulated and real-world robot experiments, VIP rewards empower robots to learn new skills without the need for manual human reward labeling. Notably, VIP demonstrates the first *few-shot* offline reinforcement learning pipeline for real-world manipulation, using just 20 sub-optimal trajectories to solve diverse tasks involving manipulating articulated, soft, transparent, and deformable objects.

Multi-Modal Value Pre-Training. While VIP can act as zero-shot visual reward functions, its dependence on image-based goals can pose challenges. This mode of specification, while efficient, might be counterintuitive for the average user and can inadvertently incorporate irrelevant aspects of a scene into the task specification. Language, on the other hand, offers a more intuitive and user-friendly goal interface. This raises the question: how can we gauge visual task progression aligned with language-specified objectives? In LIV [19], we make the surprising theoretical finding that VIP's goal-conditioned RL objective is a natural generalization to language-image contrastive learning (e.g., CLIP [28]) when the goal is specified via language. This discovery paves the way for Language-Image Value (LIV), a theoretically grounded extension of VIP that enables multi-modal goal specification. Pre-trained on text-annotated human video datasets such as EpicKitchen [7], the resulting LIV model is capable of assigning dense rewards to individual frames of unseen robot videos when the goal is specified with either image or language; see Figure 3 for an example. The generality of the LIV objective can be adapted to modalities beyond image and text and presents a versatile algorithmic blueprint for not only how to specify tasks in any modality but also how to achieve them.

Future Work

In my future work, I strive to continue developing algorithms for robot foundation reward models, with emphasis on real-robot training and deployment.

LLM Guided Real-To-Sim-To-Real Transfer. The DrEureka and Eurekaverse results suggest that LLMs combined with search can be an effective approach for automated sim-to-real transfer. However, my work has not addressed how to use foundation models to guide the transfer of vision-based manipulation tasks from simulation. In my future work, I intend to explore how we can use a few real-world demonstrations to enable foundation models to effectively construct simulation environments for manipulation tasks to enable effective transfer of generalizable policies trained in simulation.

Structured Value Pre-Training. Both VIP and LIV learn their implicit-value representations on top of RGB image inputs. Many real-world tasks, on the other hand, would benefit from additional sensing modalities, such as touch [6, 35, 36] and sound [8, 10, 11]. In a recent work [32], we have already found LIV to remain effective even when the input images consist of segmentation masks. I plan to investigate scaling up my value pre-training algorithms to directly learn latent representations on top of structured sensing inputs to better solve fine-grained real-world manipulation tasks.

Real-World Reinforcement Learning with Pre-Trained Values While I have demonstrated that VIP can accelerate online RL in simulation by supplying rewards and frozen visual representations for policies, delivering the same result in the real world must address additional challenges, such as sample inefficiency of real-world RL [39], the need for environment reset [13, 31], and incorporate additional priors, such as human demonstrations [30, 34] and offline data [2, 15, 24]. In a recent work [38], we have found VIP's value predictions to be capable of automatically discovering subgoals in demonstrations of long-horizon, multi-stage tasks. Extending this technique to segment third-person human offline data to provide intermediate subgoals and goal-conditioned rewards is a promising approach to bootstrap real-world RL.

References

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

[2] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv:2302.02948*, 2023. 2

[3] 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. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[4] Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5920–5929, 2023. 1

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1

[6] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[8] Maximilian Du, Olivia Y Lee, Suraj Nair, and Chelsea Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. *arXiv preprint arXiv:2205.14850*, 2022. 2

[9] Aleksandra Faust, Anthony Francis, and Dar Mehta. Evolving rewards to automate reinforcement learning. *arXiv preprint arXiv:1905.07628*, 2019. 1

[10] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *arXiv preprint arXiv:2109.07991*, 2021. 2

[11] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022. 2

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[13] Abhishek Gupta, Justin Yu, Tony Z Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6664–6671. IEEE, 2021. 2

[14] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023. 1

[15] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pages 1702–1712. PMLR, 2022. 2

[16] William Liang, Sam Wang, Hung-Ju Wang, Osbert Bastani, Dinesh Jayaraman*, and Yecheng Jason Ma*. Environment curriculum generation via large language models. In *Conference on Robot Learning (CoRL)*, 2024. 1

[17] Yecheng Jason Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In *International Conference on Machine Learning (ICML)*, 2022. 2

[18] Yecheng Jason Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned rl vis f-advantage regression. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2

[19] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning (ICML)*, 2023. 1, 2

[20] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual rewards and representations via value-implicit pre-training. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2

[21] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[22] Yecheng Jason Ma*, William Liang*, Hung-Ju Wang,

Sam Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. Dreureka: Language model guided sim-to-real transfer. In *Robotics: Science and Systems (RSS)*, 2024. 1

[23] Fabio Muratore, Fabio Ramos, Greg Turk, Wenhao Yu, Michael Gienger, and Jan Peters. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, page 31, 2022. 1

[24] Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *arXiv preprint arXiv:2303.05479*, 2023. 2

[25] Scott Niekum, Andrew G Barto, and Lee Spector. Genetic programming for reward function search. *IEEE Transactions on Autonomous Mental Development*, 2(2):83–90, 2010. 1

[26] OpenAI. Gpt-4 technical report, 2023. 1

[27] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018. 1

[28] 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. 2

[29] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR, 2015. 2

[30] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 2

[31] Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Autonomous reinforcement learning via subgoal curricula. *Advances in Neural Information Processing Systems*, 34:18474–18486, 2021. 2

[32] Junyao Shi, Jianing Qian, Yecheng Jason Ma, and Dinesh Jayaraman. Composing pre-trained object-centric representations for robotics from "what" and "where" foundation models. In *International Conference on Robotics and Automation (ICRA)*, 2024. 2

[33] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 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)*, pages 23–30. IEEE, 2017. 1

[34] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 2

[35] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022. 2

[36] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023. 2

[37] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023. 1

[38] Zichen Zhang, Yunshuang Li, Osbert Bastani, Abhishek Gupta, Dinesh Jayaraman, Yecheng Jason Ma*, and Luca Weihs*. Universal visual decomposer: Long-horizon manipulation made easy. In *International Conference on Robotics and Automation (ICRA)*, 2024. 2

[39] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570*, 2020. 2