Towards Efficient Generalization of Reinforcement Learning Agents by Leveraging Human and Learned Priors

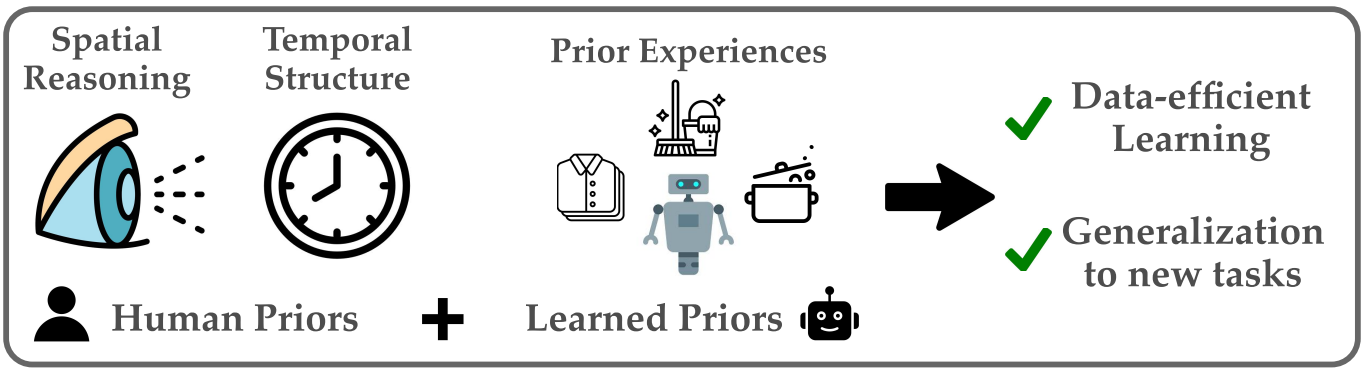

Figure 1: Humans inherently possess rich spatial-temporal reasoning capabilities and vast prior experiences that allow us to reason about and adapt to new environments and tasks. Reinforcement learning agents can benefit from these rich human priors, inductive biases, and coupled with their own prior experiences generalize to new tasks in a data-efficient manner.

1 Motivation and Research Overview

Reinforcement Learning (RL) agents have demonstrated impressive performance across many applications including video games, robotics, healthcare, and recommendation systems. Yet, they remain highly data-inefficient, often requiring millions of environment interactions or hundreds of expert demonstrations to solve simple tasks. For instance, a state-of-the-art RL agent requires playing over 200,000 games – the equivalent of a couple months of human game play – to solve the simple Atari Pong game, while humans can learn to play it in just several minutes with little to no prior experience. Moveover, RL agents often overfit to the narrow range of tasks that they were trained on, struggling to generalize to new tasks; an agent trained to play Pong will fail catastrophically on another Atari game. In contrast, humans leverage their vast prior knowledge and experiences to quickly adapt to entirely new environments and tasks. My research aims to bridge this gap between how humans and RL agents learn by **developing algorithms that leverage human priors and learned priors to bootstrap learning for decision-making agents resulting in more *data-efficient learning* and *generalist behavior***.

In my past work, I have focused on two types of human priors: *task-specific knowledge* and *temporal structure* (see Figure 1). Task-specific information includes the task itself (e.g. what is the objective?) and state representations (e.g. which parts of an image are relevant for decision-making?). This task-specific knowledge can be communicated by a human in the loop to an RL agent both explicitly through cues like natural language or expert demonstrations, and implicitly, through gesture or gaze. In one project, I used human gaze to convey important parts of a scene to an RL agent. Our gaze locations and patterns carry task-relevant information that may be hard to articulate explicitly, offering a low-effort mechanism for guiding RL agents. Based on this insight, I developed a framework using human gaze to enhance visual representations leading to more sample-efficient and robust robot policies [Liang et al., 2024c].

In addition to visual perception, humans are adept at temporal reasoning, synthesizing information across time to predict, and plan into the future. We reason about uncertainty and incorporate it to make better, more informed decisions. In a separate line of work, I investigated a sequential-decision making setting where latent factors that affect the generative process change over time. Many real-world applications exhibit this property such as recommendation systems in which the user state (e.g. mood) is unobserved and non-stationary. A recommender agent must maintain a *belief* over the user's state and model how it changes over time, creating a difficult learning problem. I proposed a meta-RL algorithm which exploits this temporal structure to learn a more accurate belief model and efficiently solve the target task [Liang et al., 2024a].

While human priors provide valuable information for RL agents to learn more effectively, they may not be sufficient to achieve agents capable of diverse behaviors. When attempting a completely new task, we typically leverage prior knowledge and experiences to bootstrap our learning. For example, when learning to play tennis, we rely on our experience with other racket sports like badminton or racquetball. In my previous work, I proposed a framework for scalable and modular skill learning [Liang et al., 2022]. First, I pretrained a large Transformer backbone model to capture diverse behaviors. For each new downstream task, the base model is frozen and a small *adapter* network (only 2% of the base model parameters) efficiently adapts to the new task by leveraging the prior knowledge encoded in the base model and a few expert demonstrations. Once trained, these task-specific adapters are modular subunits that can plug-and-play into the base model to produce different desired behaviors. Learned adapters for primitive

skills can further be combined with each other to perform compositional tasks. While we can continually accumulate adapters, each new skill requires new expert demonstrations with labelled actions. What if the pretraining dataset is *unlabelled*, and we only have a few or just one example to learn a new downstream task? In my current research, I am developing a method for learning RL agent policies from Internet videos, which do not have action labels, and employ in-context learning to generalize to a new domain with as few as a single demonstration.

Proposed Two-Year Research Thrusts:

1. Investigate other natural, readily available forms of human priors to accelerate RL training including intuitive physics and causal reasoning;
2. Develop a set of standard tasks and algorithms in simulation and real-world applications to systematically train, evaluate, and understand an RL agent's capacity to efficiently *generalize* to unseen tasks.

2 Prior Work: Human Gaze and Temporal Reasoning

Human Gaze Informed Visual Representations for Robot Learning. For robots, images are cluttered with distractions, such as background objects and varying lighting conditions, making it difficult for them to extract useful features for decision-making. Teaching robots to perform complex control tasks from high-dimensional image inputs is a nontrivial problem. State-of-the-art methods require hundreds of manually collected expert demonstrations or hours to days of training in simulation to learn a simple pick-and-place skill. RL agents must effectively compress pixel data into a lower-dimensional representation.

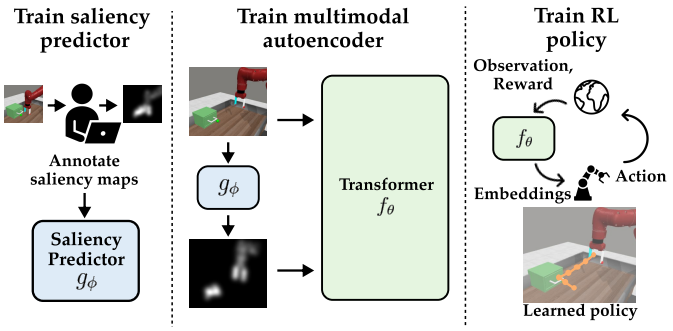

Figure 2: ViSaRL trains a saliency prediction model from a few human-annotated saliency maps. A visual encoder is pretrained with paired image and pseudo-labelled saliency data to generate image representations for the RL policy.

Studies in neuroscience found that humans utilize selective attention to focus on task-relevant information for efficiently processing complex visual scenes [Darby et al., 2021]. When performing everyday pick-and-place tasks, we employ selective attention to identify the target objects, focus on the grasp points, and execute precise hand-eye coordination. Drawing inspiration from our visual mechanisms, I developed a method to learn visual representations that capture useful features of the sensory input to simplify the decision-making process. While prior works proposed to learn such representations through various self-supervised objectives, such as contrastive learning [Laskin et al., 2020] and data augmentation [Kostrikov et al., 2021], I focus on *saliency*, a continuous heatmap that represents the importance of different parts of an image tracked by our gaze patterns. *Saliency introduces additional human domain knowledge to inform the representation of task-relevant features in the visual input and filters out uninformative perceptual noise.*

I introduced **Vi**sual **Sa**liency **R**einforcement **L**earning (ViSaRL) [Liang et al., 2024c], **a general approach for incorporating human saliency maps as an inductive bias to improve visual representations**. The key idea of ViSaRL is to train a visual encoder using both RGB and saliency inputs and an RL policy that operates over the image representations from the encoder. I trained a multimodal autoencoder using a self-supervised masked reconstruction objective and demonstrated that the learned representations attend to the most salient parts of the input image. Training the visual encoder requires a large offline dataset of paired RGB and corresponding saliency maps, which is expensive to manually annotate. To circumvent this overhead, I trained a state-of-the-art saliency predictor using a few human-annotated examples to pseudo-label all the RGB observations in the offline dataset with saliency maps. ViSaRL achieves an 18% improvement in task success rate compared to state-of-the-art baselines in simulation experiments and nearly doubles the task success rate in real-robot experiments. I also demonstrated that ViSaRL is robust to various visual perturbations, such as different backgrounds, distractor objects, and cluttered tabletops. In this work, I built on the simple intuition that human gaze provides an informative prior over the relevant parts of an image, and proposed a framework to integrate human gaze for training more robust robot policies.

Building on ViSaRL, I am currently developing a method to utilize human gaze for assisting robot teleoperation. Human gaze not only captures pertinent visual information, but also conveys our intent and preferences which is essential for developing robots that can work alongside humans in a collaborative and assistive manner.

Proposed future direction: Investigate types of implicit human feedback beyond gaze such as gestures and incorporate these complementary modalities into ViSaRL to learn better representations for decision-making.

With this framework, a human operator could simply look in the direction of the target object or location of interest to *implicitly* signal to the robot what actions it should take. I aim to leverage information from human gaze to reduce the cognitive and physical load for a human in the loop to guide a robot through complex tasks.

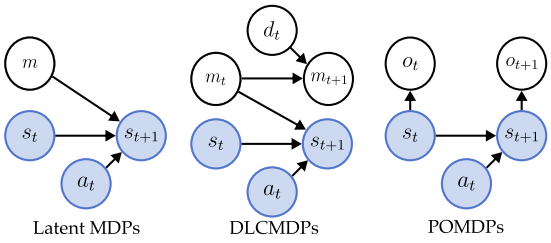

Figure 3: Dynamic Latent Context MDPs exploit the inductive bias that latent context changes infrequently according to unobserved dynamics, allowing RL agents to solve a simpler learning problem.

Leveraging temporal priors for sequential-decision making under uncertainty. Humans also excel at temporal reasoning, synthesizing past experiences and information to make informed predictions about future events. Our ability to integrate spatial and temporal cues allows us to anticipate events and adapt quickly in dynamic environments. An assumption in traditional RL approaches is that the underlying process, formally modelled as a Markov Decision Process (MDP), operates within a static environment with fixed transition dynamics and rewards [Zintgraf et al., 2019].

In many real-world systems, however, the environment dynamics are determined by latent factors that are non-stationary. For example, in a recommendation system setting, a user's preference is a latent variable which gradually evolves over time. A user may initially prefer a particular genre (e.g., action movies), but their viewing habits could change over time, influenced by external factors such as trending movies, mood, location, etc. **A robust agent should adapt to these evolving tastes to provide suitable recommendations**. During my internship project at Google Research, I formalized this setting in which latent contexts evolve at infrequent intervals as a Dynamic Latent Context MDP (DLCMDP). DLCMDPs are between latent MDPs, in which the latent context is fixed throughout the entire episode, and POMDPs, in which the latent context is always changing (see Figure 3). I proposed **Dyna**mic **M**odel for **I**mproved **T**emporal M**e**ta **R**einforcement **L**earning (DynaMITE-RL) [Liang et al., 2024b], an algorithm that exploits this temporal inductive bias to efficiently learn an RL policy by obviating the need to solve a fully general POMDP.

DynaMITE-RL consists of three crucial insights to enable tractable and efficient policy learning in DLCMDP environments. First, we consider consistency of latent information, by exploiting timesteps for which we have high confidence that the latent variable is constant. To do so, DynaMITE-RL introduces a consistency loss to regularize the posterior update model, providing better posterior estimates of the latent variable. Second, DynaMITE-RL enforces the posterior update model to learn the dynamics of the latent variable. This allows the trained policy to better infer, and adapt to, temporal shifts in latent context in unknown environments. Finally, DynaMITE-RL demonstrates that the variational objective in contemporary meta-RL algorithms, which attempts to reconstruct the entire trajectory, can hurt performance when the latent context is non-stationary. DynaMITE-RL modifies this objective to reconstruct only the transitions within the same session. I demonstrated the importance of each modification in DynaMITE-RL across various domains, ranging from discrete gridworld environments to continuous control and simulated robot assistive tasks, in both online and offline RL settings. On a challenging itch-scratching task shown in Figure 4, DynaMITE-RL achieves over 100% improvement in evaluation return over the next best baseline.

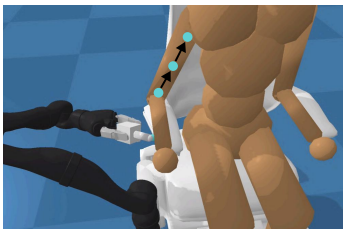

Figure 4: Assistive itching scratch task where the itch location changes over time.

Proposed future direction: Apply DynaMITE-RL in real-world applications like recommendation systems and real-robot navigation with *long-horizon contexts, non-Markovian dynamics, and visual inputs.*

3 Current Work: Meta-RL, Learning from Videos for Generalist Agents

Improving imitation learning by querying human experts on bottleneck states. In imitation learning (IL), an agent learns to mimic the actions of an expert demonstrator. Imitation learning for many real-world applications suffer from a few well-known problems: 1) many expert demonstrations are required, 2) new demonstrations can be prohibitively expensive to collect and do not guarantee improvement in the agent's performance, and 3) covariate shift can lead to degenerate behavior during test time. Expert demonstrations only cover a narrow distribution of states that the agent will encounter when deployed into the real-world. Consequently, the agent will accumulate small errors in its predictions which over many timesteps can lead to catastrophic failures.

I propose to address the shortcomings of traditional IL approaches in an *active imitation learning* framework [Judah et al., 2014] whereby the agent can acquire more data by querying a human expert for action labels. The objective is to minimize the burden on the human while maximizing the trained agent's task performance on a single deployment. Prior work asks human expert to annotate entire trajectories with action labels which is expensive and not scalable. Rather than uniformly querying for additional data along a trajectory, we can again leverage the human

prior that **certain bottleneck states are more sensitive and mission-critical, requiring more granular supervision**. For example, to open a drawer with a handle, the robot's approach to the handle is less important compared to its ability to precisely manipulate the handle once it is near it. I propose to *automatically identify critical bottleneck states* in the expert trajectory and query for a few additional steps of expert actions at these states. I design and evaluate several acquisition functions for measuring the criticality of a state including the variance of a policy ensemble, influence functions, and a Bayesian objective based on maximizing information gain. The top-K states with the highest value computed by the acquisition function are chosen to query the human for new action labels. This approach requires significantly less manual effort and ensures robustness at critical points along the trajectory. I hypothesize that with only a small amount of new data samples at informative states, we can achieve similar task performance as collecting full trajectories.

Training generalist agents using Internet-scale videos. When learning a new sport like tennis, we usually start by watching tutorial videos on YouTube. This process, known as *learning from observation* [Torabi et al., 2019], allows us to directly translate from watching videos to the physical motions of swinging a tennis racket even if we had never used one in the past. Similarly, RL agents could benefit from the vast pool of instructional videos on media platforms like YouTube to acquire new skills. Recently, Internet scale data have been used to train powerful models with diverse capabilities in language understanding and visual reasoning. **Internet videos are also easily accessible and provide data across a multitude of tasks and real-world scenes, making them a valuable data source for training generalist agents**. A key challenge in learning from video or observation-only data is the absence of action and reward labels which are necessary for any RL training. In my recent work, I developed Prompt-DTLA, a framework for training generalist agents from unlabelled videos [Liang et al., 2024a].

Given a large corpus of Internet videos spanning various tasks, the goal is to train an agent that can generalize to new environments and new tasks with very few expert demonstrations. To overcome the lack of action labels, I first trained a Latent Dynamics Model (LAM) following [Schmidt and Jiang, 2024, Bruce et al., 2024] to infer *latent actions* from observation history (see Figure 5). Then, using a small set of labelled data from any behavioral policy, I trained an action decoder that maps latent actions to ground-truth environment actions. Finally, I trained an autoregressive Transformer-based policy conditioned on few-shot expert prompts and found that it can in-context generalize to new unseen tasks during inference time. I demonstrated that unsupervised training of a LAM on the full offline dataset is more data-efficient than training an inverse dynamics model (IDM) [Baker et al., 2022] which can only learn from labelled data. Prompt-DTLA achieves over 40% improvement in success rate over the best baseline on challenging generalization tasks in XLandMiniGrid [Nikulin et al., 2023].

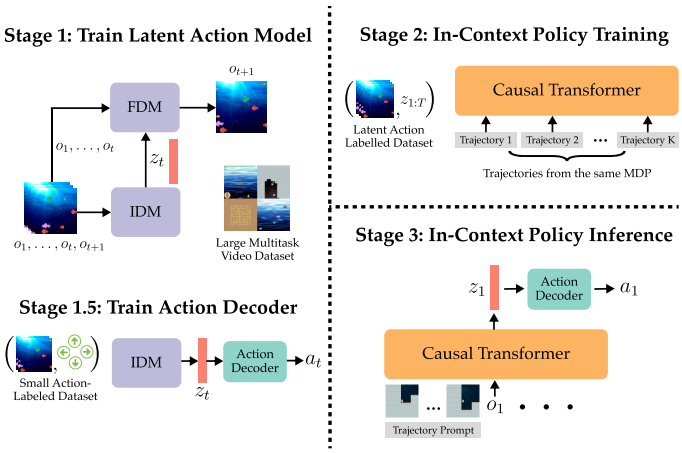

Figure 5: Prompt-DTLA is a framework for training generalist agents from videos. We first learn a Latent Action Model to infer latent actions for a large video corpus. We then train an in-context transformer policy that generalizes to new unseen tasks with few-shot examples.

Currently, I am scaling Prompt-DTLA to more challenging visual environments like Procgen [Cobbe et al., 2019], a collection of procedurally generated platformer games. In Procgen, each *level* corresponds to a different random seed used to generate the task instance, including the layout, entities and visual aspects and each *game* also has different dynamics and objective, making it a nontrivial generalization problem. I am also conducting ablation studies to understand the effects of different hyperparameters in the LAM (e.g. quantization method) and architecture choices (e.g. CNN, Transformers, State Space Models). Beyond 2D platformers, I am excited to extend our LAM to generalize to environments with different *action spaces*, including continuous actions. A continuous-LAM would enable scaling Prompt-DTLA to efficiently learn robot policies from human video data.

Humans have a remarkable ability to construct complex models of the world enabling us to predict future events and plan ahead effectively. In DynaMITE-RL, we construct a world model to quantify uncertainty in the agent's belief of the environment and in Prompt-DTLA, we learn a video dynamics model which can be used to predict how a scene will evolve over time. I am interested in constructing more explicit and accurate world models by incorporating human priors to bias the model. Humans inherently possess an intuitive understanding of physics including phenomenons like object permanence, causality, spatial awareness, which are useful priors for building a more realistic world model.

Proposed future direction: Investigate learning a latent action model for continuous action spaces and scale Prompt-DTLA to enable few-shot generalization for learning robot policies from human videos.

4 Conclusion

My past work explores how reinforcement learning (RL) agents can achieve more efficient learning and improved generalization by incorporating various human and learned priors. Specifically, by leveraging task-specific knowledge and temporal structure, I have shown that RL agents can efficiently learn to solve real-robot manipulation tasks and incorporate uncertainty to adapt in environments with changing dynamics. Moving forward, I plan to continue investigating additional forms of human priors and feedback mechanism, developing algorithms that leverage these modalities to build richer representations and robust RL agents.

References

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video Pretraining (VPT): Learning To Act By Watching Unlabeled Online Videos. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative Interactive Environments. In *International Conference on Machine Learning (ICML)*, 2024.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019.

Kevin P Darby, Sophia W Deng, Dirk B Walther, and Vladimir M Sloutsky. The development of attention to objects and scenes: From object-biased to unbiased. *Child development*, 2021.

Kshitij Judah, Alan Paul Fern, Thomas G Dietterich, and Prasad Tadepalli. Active Imitation Learning: Formal and Practical Reductions to IID learning. In *Journal of Machine Learning Research (JMLR)*, 2014.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations (ICLR)*, 2021.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2020.

Anthony Liang, Ishika Singh, Karl Pertsch, and Jesse Thomason. Transformer Adapters for Robot Learning. In *Conference on Robot Learning (CoRL) Workshop on Pre-training for Robot Learning*, 2022.

Anthony Liang, Pavel Czempin, Yutai Zhou, Stephen Tu, and Erdem Biyik. In-Context Generalization to New Tasks From Unlabeled Observation Data. In *International Conference on Machine Learning ICML 2024 Workshop on In-Context Learning*, 2024a.

Anthony Liang, Guy Tennenholtz, Chih-wei Hsu, Yinlam Chow, Erdem Bıyık, and Craig Boutilier. DynaMITE-RL: A Dynamic Model for Improved Temporal Meta-Reinforcement Learning. In *International Conference on Machine Learning (ICML) AutoRL Workshop, In submission to Conference on Neural Information Processing Systems (NeurIPS)*, 2024b.

Anthony Liang, Jesse Thomason, and Erdem Bıyık. ViSaRL: Visual Reinforcement Learning Guided by Human Saliency. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024c.

Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Viacheslav Sinii, Artem Agarkov, and Sergey Kolesnikov. XLand-minigrid: Scalable meta-reinforcement learning environments in JAX. In *Intrinsically-Motivated and Open-Ended Learning Workshop, NeurIPS2023*, 2023. URL https://openreview.net/forum?id=xALDC4aHGz.

Dominik Schmidt and Minqi Jiang. Learning to Act without Actions. In *International Conference on Learning Representations (ICLR)*, 2024.

Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representations (ICLR)*, 2019.