Fostering Intrinsic Motivation in Reinforcement Learning with Pretrained Foundation Models

Alain Andres[1]*
alain.andres@tecnalia.com

Javier Del Ser[1,2]
javier.delser@tecnalia.com

[1] TECNALIA, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastian, Spain
[2] University of the Basque Country, 48013 Bilbao, Spain

Abstract

Exploration remains a significant challenge in reinforcement learning, especially in environments where extrinsic rewards are sparse or non-existent. The recent rise of foundation models, such as CLIP, offers an opportunity to leverage pretrained, semantically rich embeddings that encapsulate broad and reusable knowledge. In this work we explore the potential of these foundation models not just to drive exploration, but also to analyze the critical role of the episodic novelty term in enhancing exploration effectiveness of the agent. We also investigate whether providing the intrinsic module with complete state information – rather than just partial observations – can improve exploration, despite the difficulties in handling small variations within large state spaces. Our experiments in the MiniGrid domain reveal that intrinsic modules can effectively utilize full state information, significantly increasing sample efficiency while learning an optimal policy. Moreover, we show that the embeddings provided by foundation models are sometimes even better than those constructed by the agent during training, further accelerating the learning process, especially when coupled with the episodic novelty term to enhance exploration.

1 Introduction

Exploration is a fundamental challenge in reinforcement learning (RL), particularly in environments where extrinsic rewards are sparse or absent. In such scenarios, agents often struggle to efficiently explore and learn, frequently failing to discover meaningful strategies that lead to successful task completion. To address these challenging RL tasks, various methods have been proposed, including approaches that maximize entropy Haarnoja et al. [2018], techniques that increase the uncertainty of the agent's parameters Fortunato et al. [2018], and imitation learning to mimic expert behaviors Zheng et al. [2024]. Albeit effective, these proposals have their own drawbacks, such as the need for careful configuration tuning, or the requirement of externally gathered data.

Inspired by human psychology, intrinsic motivation encourages agents to be curious (i.e., seek novel situations) in the environment, fostering a more robust and adaptive learning process Aubret et al. [2019]. This typically involves generating an intrinsic reward r_t^i that quantifies the novelty or *surprise*, which is then added to the actual extrinsic reward from the environment r_t^e. This addition leads to an updated reward signal $r_t = r_t^e + \beta \cdot r_t^i$, wherein $\beta \in \mathbb{R}^+$ balances between exploiting known information in the environment with fostering the agent's curiosity. However, the effectiveness of these methods often depends on the quality of the state representations used to compute the novelty bonus, which are commonly learned during the agent's training process.

---

*Corresponding author: https://aklein1995.github.io

Recently, foundation models like CLIP (Contrastive Language-Image Pretraining, Radford et al. [2021]) have gained prominence in the literature for their ability to encode rich, semantically meaningful representations of visual and textual data. These models, pretrained on vast datasets, encapsulate a wide range of reusable knowledge that can be transferred across different tasks and domains. By leveraging these pretrained embeddings, agents can generate more informative intrinsic rewards, driving exploration in a more targeted and meaningful way.

This work builds on FoMoRL (Foundation Models for Semantic Novelty in Reinforcement Learning Gupta et al. [2022]), which demonstrated the effectiveness of using CLIP embeddings to drive exploration. Our research takes a step further by posing and answering with empirical evidence two main research questions:

- How effective are foundation models in generating embeddings that enhance exploration in sparse RL? Can their capabilities be extended by providing full state information?

  Previous studies have highlighted the challenges of computing novelty from full states Campero et al. [2021], where methods like Random Network Distillation (RND) Burda et al. [2018] and Counts Bellemare et al. [2016] struggle to differentiate states within the same episode (yet increase surprise between episodes). While FoMoRL has shown promise with full state information, it remains uncertain whether more sophisticated techniques like RIDE Raileanu and Rocktäschel [2020] would also perform competitively in this setup.

- Is the success of these models driven by the inclusion of an episodic novelty term, which penalizes revisiting states within an episode?

  The episodic novelty term has recently been shown to be one of the most important components when generating intrinsic rewards Andres et al. [2022], Wang et al. [2023]. In FoMoRL, it is unclear whether such term was in use, and if so, how it was computed from the RGB images collected by the agent. Decoupling the contribution of foundational model embeddings from the episodic novelty term is crucial for understanding their respective roles.

To address these questions, we conduct a series of experiments over the MiniGrid domain, comparing the performance of agents using different configurations of RIDE and the FoMoRL approach. Throughout this investigation, we aim to provide deeper insights into the mechanisms that drive successful exploration in RL, and explore the potential of foundation models to push the boundaries of open-ended learning.

2 Related Work

Intrinsic Motivation for enhanced exploration in RL. In response to the novelty that the agent experiences during its learning process, the reward signal can be augmented to foster explorative behaviors. This is usually tackled via count-based methods that encourage visits to less frequent states Bellemare et al. [2016], Ostrovski et al. [2017], or by reducing uncertainty/prediction error in the environment where the agent is deployed Pathak et al. [2017], Burda et al. [2018]. These *lifelong* rewards are significantly beneficial when being combined with *episodic* information Raileanu and Rocktäschel [2020], Flet-Berliac et al. [2021], Zhang et al. [2022], being sometimes even more interesting to use solely the latter episodic information Andres et al. [2022], Henaff et al. [2023a], Wang et al. [2023]. Thus, deriving methods that capture this episodic information has become of interest within the community working on sparse RL tasks Savinov et al. [2019], Jo et al. [2022], Henaff et al. [2023b].

Foundational Models and RL. Recently, some approaches have relied on Vision Language Models (VLM) to generate auxiliary rewards by calculating the cosine similarity between the language goal and the input image Mahmoudieh et al. [2022]. However, their effectiveness seems to be related to the model size and the dataset used for learning them Rocamonde et al. [2024], Baumli et al. [2024]. While these rewards can be noisy, they have proven to be useful during pretraining Adeniji et al. [2023] and can be refined through contrastive reward alignment techniques Fu et al. [2024]. Similarly, VLM embeddings have been employed to generate intrinsic rewards, driving exploration more effectively in sparse environments Gupta et al. [2022]. Additionally, LLMs have been leveraged to guide the learning process, providing high-level instructions Du et al. [2023] and shaping the reward function Chu et al. [2023].

3 Methodology

In this section we motivate the research questions stated in the introduction based on the definition of the rewards at the core of RIDE ad FoMoRL.

RIDE: Rewarding Impact-Driven Exploration. RIDE is one the state-of-the-art exploration strategies proposed for environments with sparse rewards. It computes intrinsic rewards based on the impact an agent's actions have on the environment. Specifically, it measures the difference between consecutive states in a learned state representation space, with larger differences indicating more significant exploration. Mathematically:

$$r_t^{RIDE}(s_t, s_{t+1}, a_t) = \frac{\|\phi(s_{t+1}) - \phi(s_t)\|_2}{\sqrt{N_{\text{ep}}(s_{t+1})}}, \tag{1}$$

where $\phi(\cdot)$ refers to the embedding network that is learned during training, $\|\cdot\|_2$ stands for L_2-norm, and $N_{ep}(s_{t+1})$ is the number of times the next state s_{t+1} has been visited.

FoMoRL: Foundation Models for Semantic Embeddings. FoMoRL proposes a different approach by using pretrained state representations from a foundation model, instead of relying on domain-specific state representations that need to be learned, thereby accelerating exploration. The intrinsic reward in FoMoRL is computed as:

$$r_t^{FoMoRL}(s_t, s_{t+1}, a_t) = \frac{\|\text{clip}(s_{t+1}) - \text{clip}(s_t)\|_2}{\sqrt{N_{\text{ep}}(s_{t+1})}}, \tag{2}$$

where $\phi(\cdot)$ is replaced by the outputs of CLIP's visual embeddings.

Motivation. Although Equations (1) and (2) were originally proposed considering the full information about the environment, in practice they have been applied using either the state s_t or the partial observation o_t. Furthermore, depending on the environment, this information can be represented as an encoded/compact representation, $\{s_t^{\text{enc}}, o_t^{\text{enc}}\}$, or as an RGB visualization when images are used, $\{s_t^{\text{rgb}}, o_t^{\text{rgb}}\}$. Our work focuses on understanding the impact of these different representations on exploration. Specifically, we aim to analyze: (1) the effect of using either state or observation in the intrinsic curiosity module, (2) the differences between employing an encoded representation or an image, and (3) the importance of the episodic novelty term in these diverse setups.

4 Experimental Setup

MiniGrid. Experiments are conducted over the MiniGrid benchmark, a collection of procedurally generated grid-based environments that challenge RL agents to explore and navigate effectively in the presence of reward sparsity. In MiniGrid, the agent can receive environment information in one of two ways: as a full state, where the agent has complete access to the entire grid layout, s_t, or as a partial observation, o_t, where the agent's view is restricted to a limited, egocentric field centered on its current position. Partial observations are thus a subset of the full state, unable to capture the areas behind walls or closed doors. The information, whether full state or partial observation, can be represented in two forms: as a compact encoded vector (e.g., s_t^{enc}, o_t^{enc}) that summarizes the environment's key elements, or as an RGB image (e.g., s_t^{rgb}, o_t^{rgb}).

Figure 1 illustrates these different observation types, showing examples of both encoded and RGB representations, as well as full and partial observations. We evaluate performance across several MiniGrid environments, including MultiRoom (MN7S4, MN7S8, MN12S10), KeyCorridor (KS3R3, KS4R3), and ObstructedMaze (O2Dlh).

Algorithms. For training an agent, we use Proximal Policy Optimization (PPO) Schulman et al. [2017] across all experiments, utilizing RIDE and FoMoRL to generate intrinsic rewards. In our setup, RIDE operates exclusively with encoded information, using either the encoded state s_t^{enc} or observation o_t^{enc} to compute the intrinsic rewards. Conversely FoMoRL, which is pretrained on image data, utilizes RGB representations s_t^{rgb} or o_t^{rgb} for the calculation of its reward. Regardless of the intrinsic motivation technique, when incorporating the episodic novelty term, we use the encoded observation $N_{ep}(o_t^{enc})$ to track state visitation counts. Similarly, the agent's policy is always fed with encoded observations $\pi(\cdot|o_t^{enc})$. More information about the selected model architectures and hyperparameters can be found in Appendix A.
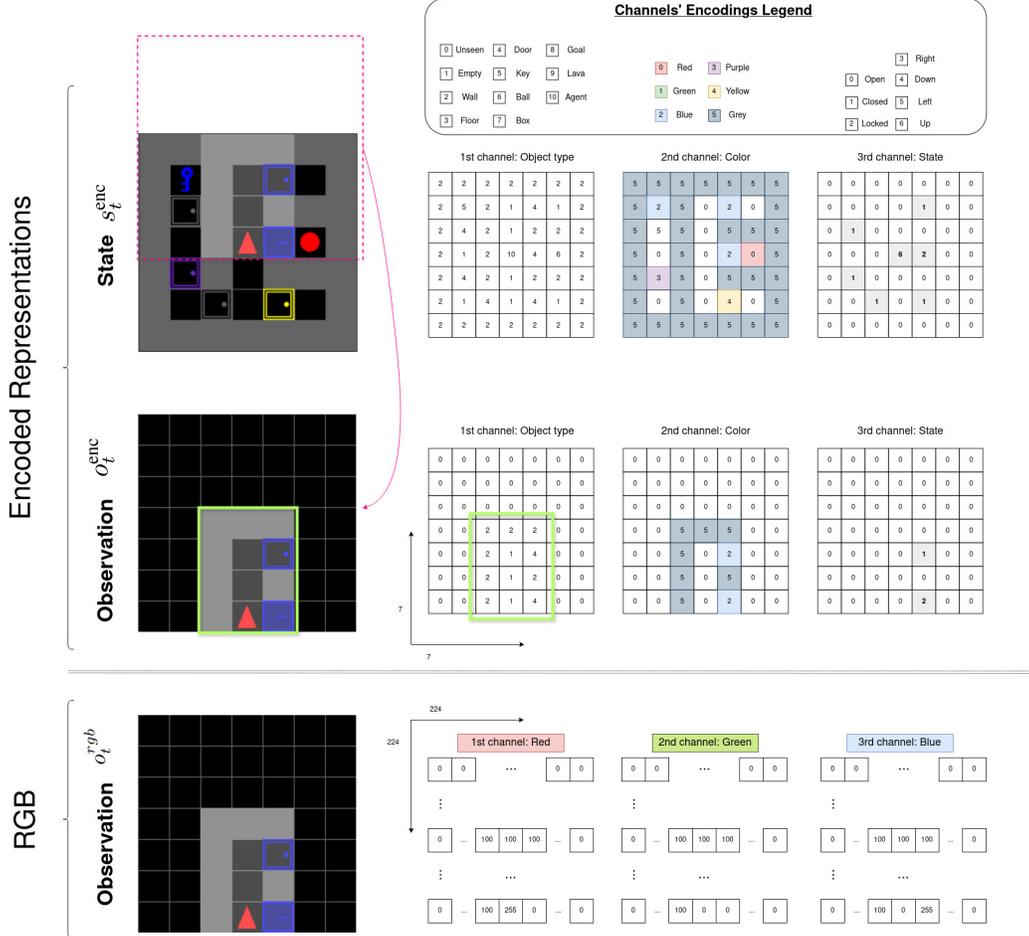
Figure 1: Illustration of full state and partial observation formats in the MiniGrid environment, with two types of representations: encoded and RGB. The top row shows the full state s_t^{enc}, where the agent has access to the entire grid layout, represented as a three-channel encoded matrix. The three channels encode object type, color, and state (e.g., open or closed doors, agent orientation). The dotted pink square in this row indicates the region that would compose the agent's observation in a partial view. The middle row displays this partial observation o_t^{enc}, where the agent's perception is limited to a 7×7 egocentric field centered on its position. The green square in this row highlights the area visible to the agent, showing how walls and other objects limit the agent's field of vision. Lastly, the bottom row shows the RGB representation of a partial observation o_t^{rgb}, where each cell is represented according to three color channels (red, green, and blue) for a pixel-based view.

5   Results

Impact of Partial vs Full Observability.   Table 1 provides the number of steps required for RIDE and FoMoRL to converge to the optimal policy in different MiniGrid environments, considering both partial observations and full state information. For both approaches, *the use of full state information significantly augments the sample-efficiency*. In complex `MultiRoom` environments only FoMoRL with full state access manages to effectively train. However, in `KeyCorridor`, RIDE performs slightly better than FoMoRL. We hypothesize that this is because these environments contain a higher diversity of objects, allowing RIDE to effectively distinguish between different states and learn a better representation. In contrast, FoMoRL relies on pretrained embeddings from CLIP, which remain frozen during training and may lack the adaptability required to fully capture the variations in these more diverse environments.

Table 1: Performance of RIDE and FoMoRL in terms of the number of required steps to achieve convergence in various MiniGrid environments after training on [xM] steps. The (-) indicates that the learning was unsuccessful. The first two columns (Partial) consider only observations for RIDE (o_t^{enc}) and FoMoRL (o_t^{rgb}), while the last two columns (Full) consider state information for RIDE (s_t^{enc}) and FoMoRL (s_t^{rgb}). All the results make use of the episodic term component, $\sqrt{N_{ep}(o_t^{enc})}$.

	Partial, o_t		Full, s_t	
	RIDE	FoMoRL	RIDE	FoMoRL
`MultiRoom-N7-S4` [10M]	**1.53M**	4.03M	0.93M	**0.58M**
`MultiRoom-N7-S8` [20M]	-	-	-	**2.8M**
`MultiRoom-N12-S10` [30M]	-	-	-	**4.06M**
`KeyCorridorS3R3` [30M]	6.82M	**5.29M**	**2.07M**	5.13M
`KeyCorridorS4R3` [50M]	-	**17.06M**	**7.92M**	8.15M
`ObstructedMaze-2Dlh` [50M]	39.03M	**19.92M**	29.02M	**16.53M**

Role of the Episodic Term. Table 2 highlights the critical importance of the episodic term in accelerating the learning process. Without this term, the sample efficiency is significantly reduced. In fact, the absence of the episodic term has a particularly negative impact on RIDE's performance, as it is unable to learn in `MultiRoom-N7-S8` and `MultiRoom-N12-S10`. FoMoRL, however, demonstrates greater resilience to the removal of the episodic term, likely due to its reliance on latent knowledge from pretrained embeddings. Nonetheless, akin to the results with the episodic term, FoMoRL still obtains slightly worse results in some environments.

Table 2: Performance of RIDE and FoMoRL in terms of the number of required steps to achieve convergence, without considering the episodic term. That is, $\forall o_t \rightarrow \sqrt{N_{ep}(o_t)} = 1$ in Expressions (1) and (2).

	RIDE (Full, s_t^{enc})	FoMoRL (Full, s_t^{rgb})
`MultiRoom-N7-S4` [20M]	**3.19M**	14.92M
`MultiRoom-N7-S8` [30M]	-	**11.89M**
`MultiRoom-N12-S10` [50M]	-	**25.32M**
`KeyCorridorS3R3` [50M]	**5.27M**	7.12M
`KeyCorridorS4R3` [50M]	**11.95M**	-
`ObstructedMaze-2Dlh` [80M]	24.76M	**15.54M**

We refer to Appendix B for more details about the results summarized in Tables 1 and 2.

6 Conclusions

We have investigated the use of foundation models and episodic novelty to enhance exploration in sparse RL environments. Our results evince that providing full state information to the intrinsic module significantly accelerates the agent's learning process, enabling faster convergence to learn optimal policies. The episodic novelty term is critical for effective exploration; without it, agents in some environments struggle to learn. However, when prelearned state embeddings are used (as in FoMoRL), access to full state information can reduce the dependence on the episodic term. This benefit is lessened if the embedding must be learned during training (as in RIDE), as the initial exploration can hinder the characterization of suitable state representations.

In the future we plan to extend our approach to more MiniGrid environments, other RL benchmarks with visual inputs, and explore alternatives such as alignment approaches that build on top of the embeddings provided by CLIP and other foundation models.

References

Ademi Adeniji, Amber Xie, Carmelo Sferrazza, Younggyo Seo, Stephen James, and Pieter Abbeel. Language Reward Modulation for Pretraining Reinforcement Learning, August 2023. URL `http://arxiv.org/abs/2308.12270`. arXiv:2308.12270 [cs].

Alain Andres, Esther Villar-Rodriguez, and Javier Del Ser. An Evaluation Study of Intrinsic Motivation Techniques Applied to Reinforcement Learning over Hard Exploration Environments. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, pages 201–220, Cham, 2022. Springer International Publishing. ISBN 978-3-031-14463-9. doi: 10.1007/978-3-031-14463-9_13.

Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning, August 2019. URL `http://arxiv.org/abs/1908.06976`. arXiv:1908.06976.

Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, Clare Lyle, Hussain Masoom, Kay McKinney, Volodymyr Mnih, Alexander Neitz, Dmitry Nikulin, Fabio Pardo, Jack Parker-Holder, John Quan, Tim Rocktäschel, Himanshu Sahni, Tom Schaul, Yannick Schroecker, Stephen Spencer, Richie Steigerwald, Luyu Wang, and Lei Zhang. Vision-Language Models as a Source of Rewards, July 2024. URL `http://arxiv.org/abs/2312.09187`. arXiv:2312.09187 [cs].

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, Barcelona, November 2016. arXiv. doi: 10.48550/arXiv.1606.01868. URL `http://arxiv.org/abs/1606.01868`.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by Random Network Distillation, October 2018. URL `http://arxiv.org/abs/1810.12894`. arXiv:1810.12894.

Andres Campero, Roberta Raileanu, Heinrich Küttler, Joshua B. Tenenbaum, Tim Rocktäschel, and Edward Grefenstette. Learning with AMIGo: Adversarially Motivated Intrinsic Goals, February 2021. URL `http://arxiv.org/abs/2006.12122`. arXiv:2006.12122 [cs, stat].

Kun Chu, Xufeng Zhao, Cornelius Weber, Mengdi Li, and Stefan Wermter. Accelerating Reinforcement Learning of Robotic Manipulations via Feedback from Large Language Models, November 2023. URL `http://arxiv.org/abs/2311.02379`. arXiv:2311.02379 [cs].

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding Pretraining in Reinforcement Learning with Large Language Models, September 2023. URL `http://arxiv.org/abs/2302.06692`. arXiv:2302.06692 [cs].

Yannis Flet-Berliac, Johan Ferret, Olivier Pietquin, Philippe Preux, and Matthieu Geist. Adversarially Guided Actor-Critic. In *International Conference on Learning Representations (ICLR)*. arXiv, February 2021. doi: 10.48550/arXiv.2102.04376. URL `http://arxiv.org/abs/2102.04376`. arXiv:2102.04376 [cs, stat].

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy Networks For Exploration. February 2018. URL `https://openreview.net/forum?id=rywHCPkAW`.

Yuwei Fu, Haichao Zhang, Di Wu, Wei Xu, and Benoit Boulet. FuRL: Visual-Language Models as Fuzzy Rewards for Reinforcement Learning, June 2024. URL `http://arxiv.org/abs/2406.00645`. arXiv:2406.00645 [cs].

Tarun Gupta, Peter Karkus, Tong Che, Danfei Xu, and Marco Pavone. Foundation Models for Semantic Novelty in Reinforcement Learning, November 2022. URL `http://arxiv.org/abs/2211.04878`. arXiv:2211.04878 [cs].

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *35th International Conference on Machine Learning (ICML)*. arXiv, August 2018. doi: 10.48550/arXiv.1801.01290. URL `http://arxiv.org/abs/1801.01290`.

Mikael Henaff, Minqi Jiang, and Roberta Raileanu. A Study of Global and Episodic Bonuses for Exploration in Contextual MDPs, June 2023a. URL http://arxiv.org/abs/2306.03236. arXiv:2306.03236 [cs].

Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via Elliptical Episodic Bonuses, January 2023b. URL http://arxiv.org/abs/2210.05805. arXiv:2210.05805 [cs].

Daejin Jo, Sungwoong Kim, Daniel Wontae Nam, Taehwan Kwon, Seungeun Rho, Jongmin Kim, and Donghoon Lee. LECO: Learnable Episodic Count for Task-Specific Intrinsic Reward, October 2022. URL http://arxiv.org/abs/2210.05409. arXiv:2210.05409 [cs].

Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-Shot Reward Specification via Grounded Natural Language. In *International Conference on Machine Learning*, 2022.

Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-Based Exploration with Neural Density Models. In *34th International Conference on Machine Learning (ICML)*. arXiv, June 2017. doi: 10.48550/arXiv.1703.01310. URL http://arxiv.org/abs/1703.01310.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *34th International Conference on Machine Learning (ICML)*. PMLR, May 2017. doi: 10.48550/arXiv.1705.05363. URL http://arxiv.org/abs/1705.05363.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Roberta Raileanu and Tim Rocktäschel. RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments. In *International Conference on Learning Representations (ICLR)*. arXiv, February 2020. doi: 10.48550/arXiv.2002.12292. URL http://arxiv.org/abs/2002.12292.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning, March 2024. URL http://arxiv.org/abs/2310.12921. arXiv:2310.12921 [cs].

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic Curiosity through Reachability. In *International Conference on Learning Representations (ICLR)*. arXiv, August 2019. doi: 10.48550/arXiv.1810.02274. URL http://arxiv.org/abs/1810.02274.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347.

Kaixin Wang, Kuangqi Zhou, Bingyi Kang, Jiashi Feng, and Shuicheng Yan. Revisiting Intrinsic Reward for Exploration in Procedurally Generated Environments. In *International Conference on Learning Representations (ICLR)*, 2023.

Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E. Gonzalez, and Yuandong Tian. NovelD: A Simple yet Effective Exploration Criterion. In *Advances in Neural Information Processing Systems (NeurIPS)*, January 2022. URL https://openreview.net/forum?id=CYUzpnOkFJp.

Boyuan Zheng, Sunny Verma, Jianlong Zhou, Ivor W. Tsang, and Fang Chen. Imitation Learning: Progress, Taxonomies and Challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6322–6337, May 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3213246. URL https://ieeexplore.ieee.org/abstract/document/9927439. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

A   Model Architectures and Hyperparameters

Intrinsic motivation. We conducted a grid search over the range $\beta \in \{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ to determine the best scaling factor for r_t^i in each environment. Table 3 summarizes the selected coefficients when considering both full and partial information with the inclusion of the episodic term. Table 4 provides the chosen values when the episodic component is not used.

Table 3: β coefficient with episodic counts.

	Partial		Full	
	RIDE	FoMoRL	RIDE	FoMoRL
MultiRoom-N7-S4	0.005	0.0005	0.05	0.005
MultiRoom-N7-S8	0.05	0.005	0.05	0.005
MultiRoom-N12-S10	0.01	0.01	0.01	0.01
KeyCorridorS3R3	0.005	0.0005	0.05	0.005
KeyCorridorS4R3	0.1	0.0005	0.1	0.001
ObstructedMaze-2Dlh	0.005	0.0005	0.005	0.0005

Table 4: β coefficient without episodic counts, i.e., $\forall o_t \rightarrow \sqrt{N_{ep}(o_t)} = 1$ in Equations (1) and (2).

	Full	
	RIDE	FoMoRL
MultiRoom-N7-S4	0.05	0.005
MultiRoom-N7-S8	0.05	0.005
MultiRoom-N12-S10	0.01	0.01
KeyCorridorS3R3	0.05	0.0005
KeyCorridorS4R3	0.1	0.0005
ObstructedMaze-2Dlh	0.005	0.0005

PPO. We used a discount factor $\gamma = 0.99$, a clipping factor $\epsilon = 0.2$, 4 epochs per training step, and $\lambda = 0.95$ for GAE. We employed 16 parallel environments to collect rolloutsS of size 128, resulting in a total horizon of 2,048 steps between updates.

Neural Network Architecture. We used an actor-critic architecture with shared weights between the actor and critic networks. The shared network consisted of three convolutional layers (each with 32 filters of size 3×3, a stride of 2, and padding of 1), followed by a fully connected layer with 256 units. This shared representation was then fed into two separate heads: one for the actor (policy) and another for the critic (value function).

B Graphical Results

(a) MultiRoom-N7-S4

(b) MultiRoom-N7-S8

(c) MultiRoom-N12-S10

(d) ObstructedMaze-2Dlh

(e) KeyCorridorS3R3

(f) KeyCorridorS4R3
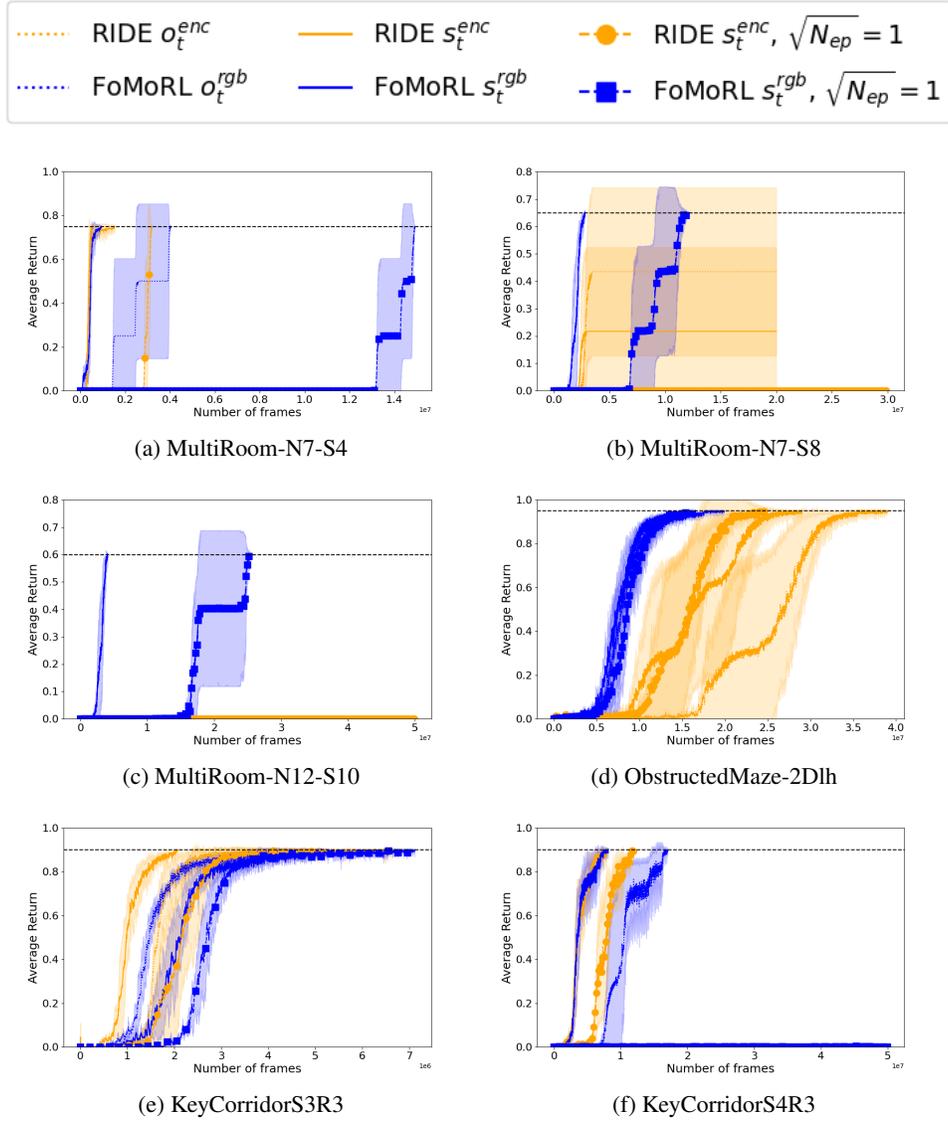
Figure 2: Performance comparison between RIDE (orange) and FoMoRL (blue) across various MiniGrid environments and input types. Each subfigure presents the average return over training steps, illustrating learning progress and convergence speed for each algorithm in specific settings. Solid lines represent full observations (s_t), while dotted lines denote partial observations (o_t). The horizontal black dashed line indicates the expected return of the optimal policy. The shaded area represents the standard deviation computed across 3 different seeds.