Mixture of Autoencoder Experts Guidance using Unlabeled and Incomplete Data for Exploration in Reinforcement Learning

1st Elias Malomgré
IDLab, Ghent University - imec
Ghent, Belgium
elias.malomgre@ugent.be

2nd Pieter Simoens
IDLab, Ghent University - imec
Ghent, Belgium
pieter.simoens@ugent.be

***Abstract*—Large amounts of weak expert-relevant data exist in practice, but current reinforcement learning methods often struggle to use it because it lacks actions, rewards, and complete trajectories. We study exploration from such weak observations and propose a framework that decouples reward construction: an expert-similarity model is learned from sparse state-only observations, and a separate white-box mapping converts expert consistency into an intrinsic reward. In our instantiation, the expert-similarity model is a mixture of autoencoder experts, and the learned signal is applied to successor states with time-dependent scaling, yielding a transient exploration prior toward expert-supported regions. We evaluate the method with Soft Actor-Critic on MuJoCo locomotion using sparse, temporally subsampled state-only observations as a proxy for sparse teleoperation data. The same learned expert-similarity model and mapping family are reused across settings, with only limited mapping-parameter adjustment. Across strong and imperfect demonstrations, the method improves exploration most clearly when extrinsic reward is sparse, partial, or otherwise insufficient for efficient exploration. These results suggest that weak expert observations can already provide useful guidance that is reusable across different reward structures, under substantially weaker assumptions than those required by standard demonstration-based policy or reward-learning methods.**

***Index Terms*—reinforcement learning, exploration, intrinsic motivation, weak supervision, reward shaping.**

I. Introduction

Exploration is central to Reinforcement Learning (RL) [1] and robotics: an agent must decide how to allocate effort under uncertainty while avoiding uninformative or undesirable parts of the state space. In RL, this is often addressed through intrinsic rewards; in robotics, through priors that bias behavior toward feasible or safer regions. Despite their differences, these formulations play a similar role: they provide evaluative signals that shape exploration. At the same time, large amounts of weak expert-relevant data already exist in the world, including teleoperation logs, traffic traces, partially observed sensor streams, and internet video [2]. Such data often contains useful information about where competent, safe, or task-relevant behavior tends to occur, even when it does not provide actions, rewards, or complete trajectories. Yet much of this data remains difficult for current RL methods to leverage directly, because it lacks the stronger structure that standard imitation, reward-learning, and replay-based pipelines typically require.

Our focus is therefore on the setting where expert-relevant data exists, but only in weak form, which current methods struggle to leverage. Demonstration-based RL typically uses expert data either as direct action supervision [3]–[5] or as additional transitions in replay buffers [5], [6]. Inverse Reinforcement Learning (IRL) [7]–[9] is closer in spirit, since it also seeks to explain expert behavior through a learned reward, but it typically does so either under stronger data assumptions or through environment and/or policy interactions, making it harder to be a reusable artifact [10]. Likewise, robotics-from-video and related passive-observation methods aim to extract control-relevant structure from weak data, but usually rely on stronger assumptions such as temporal coherence, task alignment, correspondence, or action grounding [2], [11]. By contrast, we ask a simpler but practically important question: can weak expert observations already be turned into an exploration prior, without requiring actions, full trajectories, auxiliary transition data, or online interaction during artifact construction?

This question is especially important for exploration. Generic intrinsic motivation methods [12], such as count-based exploration [13], curiosity [14], or random network distillation [15], make almost no assumptions about expert data because they do not use it at all. That generality is attractive, but it also discards potentially useful prior information about where competent behavior tends to occur. Our setting lies between these extremes. We assume expert observations that may be *unlabeled*, *incomplete*, and *imperfect*: they may lack actions, rewards, and transitions; trajectories may be fragmented or temporally sparse; and the data may be noisy, heterogeneous, or only loosely aligned with the downstream task. Rather than treating such data as direct supervision, we use it as weak behavioral evidence from which to learn an evaluative signal that guides exploration.

We propose a framework for exploration from weak expert observations through a decoupled two-component reward con-

Flanders AI Research Program

struction. First, we learn an expert-similarity model E from weak expert observations. Second, we apply a separate mapping function $g(E(s), s)$ that converts the model output into a reward. This separation is central: E captures the structure induced by weak expert data, while g determines how that structure is translated into an exploration signal. In our current instantiation, E is a mixture of autoencoder experts, and g is a white-box mapping based on reconstruction consistency. The resulting intrinsic reward is applied to successor states, yielding a transition-sensitive shaping signal that biases the agent toward regions supported by the expert. More broadly, a natural extension of the same weak-data perspective is to richer weak-observation settings, including robotics-from-video: rather than requiring weak observations to recover actions, transferable skills, or a task-complete reward [2], the same evaluative formulation could instead use them to induce a signal that biases behavior.

This formulation also gives an explicit handle on the exploration–exploitation dilemma. The learned intrinsic signal is combined with the environment reward and scaled by a time-dependent coefficient. Early in training, stronger weighting biases the agent toward expert-supported regions, improving guidance when learning from scratch is inefficient or risky. Unlike generic intrinsic motivation methods, which typically reward novelty, uncertainty, or prediction error, our signal biases exploration toward regions that are not merely new but also supported by weak expert evidence. Later in training, this influence is reduced, allowing the agent to rely more on extrinsic reward and improve beyond imperfect expert data. In this sense, weak expert data acts as a transient exploration prior rather than a fixed target behavior. The same mechanism also induces a support-based bias that can make exploration more conservative when the weak data reflect safe behavior. While we do not impose formal safety constraints or provide safety guarantees, the learned signal biases the agent toward regions represented in the expert data and away from poorly supported parts of the state space.

A second advantage of the approach is that the learned expert-similarity model and the deployed reward are distinct objects. Once E has been learned, the reward can still be reshaped through the mapping function $g(E(s), s)$ without relearning the underlying expert-similarity model. In our current instantiation, the mapping is a white-box transformation from reconstruction error to reward, making shaping behavior explicit and editable after expert-similarity modeling rather than entangling representation learning and reward design in a single end-to-end artifact. This separation suggests a broader design space in which weak-data-trained evaluators can be reused in different ways [10].

We instantiate this idea with Mixture of Expert Guidance using Unlabeled Incomplete Data for Exploration (MoE-GUIDE), here trained on sparse, temporally subsampled state-only expert observations as a proxy for sparse teleoperation data. We evaluate the method on locomotion tasks in intrinsic-only, sparse, and dense-reward settings, all using the same expert model, and find that tuning the mapping function can benefit learning. Although the present paper studies the state-based setting, the broader motivation remains the same: many real systems already contain weak observational data [2], [16]–[19] that current RL methods struggle to use directly, and enabling exploration methods to leverage such data under weaker assumptions is an important step toward more scalable and practical forms of guidance. More broadly, natural extensions of the same evaluative artifact include its use as a reusable prior in robotics or as a scoring function in search-based systems. In this sense, the broader significance of the approach is that it begins to make a class of weak expert data that current RL methods often leave unused for exploration.

Our main contributions are:

- We propose a framework for exploration from weak expert observations, where data may be static, sparse, state-only, incomplete, imperfect, and lack actions, rewards, transitions, or complete trajectories.
- We introduce a decoupled reward construction in which an expert-similarity model E captures weak-data structure and a separate white-box mapping $g(E(s), s)$ converts it into a successor-state reward.
- We instantiate this framework with MoE-GUIDE and show useful exploration gains on locomotion in intrinsic-only, dense, and sparse-reward settings.

II. Background and Related Work

We consider a Markov decision process $(\mathcal{S}, \mathcal{A}, P, r_{\text{env}}, \gamma)$, where a policy $\pi(a \mid s)$ is optimized to maximize expected discounted return. Exploration is often encouraged by augmenting the environment reward with an intrinsic term,

$$r(s, a, s', t) = r_{\text{env}}(s, a, s') + r_{\text{int}}(s, a, s', t). \tag{1}$$

In this work, the intrinsic signal is learned from weak expert observations and used to bias exploration.

We assume access to a static dataset of expert-relevant observations $\mathcal{D}_E = \{s_i\}_{i=1}^N$. In our setting, these are sparse, state-only observations sampled from expert behavior. The data is *unlabeled* in that actions, rewards, and task annotations are unavailable; *incomplete* in that transitions or full trajectories need not be observed; and *imperfect* in that observations may be noisy, heterogeneous, temporally sparse, or only loosely aligned with the downstream task. The main distinction between our setting and prior work is therefore not whether expert data is used at all, but how much additional structure is required to make that data useful.

Methods that incorporate demonstrations in RL commonly do so through BC losses [4], [5] or by inserting demonstrations into replay buffers [5], [6], thereby using expert data as direct policy supervision or privileged experience. IRL and reward-learning methods are closer in spirit because they also seek to construct a reward from static expert data, but they typically assume richer sequential structures, such as near perfect demonstrations, trajectories, transitions, or occupancy information [20]–[26]. Even state-only and state-marginal variants [25], [27] usually still require meaningful coverage,

auxiliary transition data, or environment/policy interaction during reward construction.

Methods based on passive observations, including robotics-from-video and learning from observation, also aim to extract control-relevant structure from weak data, but typically rely on additional assumptions such as temporal coherence, embodiment correspondence, task alignment, or action grounding [2], [11]. In contrast, we do not require action labels, full trajectories, explicit correspondence, or online interaction during reward construction. Instead, we learn a decoupled evaluative signal directly from weak expert observations and use it as an auxiliary reward for exploration. This shifts the role of expert data from providing direct supervision or a task-complete objective to providing partial evidence about desirable regions of the state space.

III. Method

We study exploration in reinforcement learning when weak expert observations are available as a static dataset but do not provide actions, rewards, or complete trajectories. Rather than recovering a task-complete reward from these observations, we learn an expert-similarity model from weak data and use a separate mapping function to convert the learned structure into an auxiliary reward for exploration.

A. expert-similarity model and mapping function

We decompose the method into two parts: an *expert-similarity model* $E_\theta(s)$, learned from weak expert data, and a *mapping function* $g(E_\theta(s), s)$, which converts the expert-model output and the current state into a scalar reward. This separation is central. The expert-similarity model captures structure from $\mathcal{D}_E$, while the mapping function determines how that structure is operationalized as reward. In particular, $E_\theta(s)$ does not itself define the intrinsic reward; reward is produced only after applying g.

In the present paper, g is a white-box mapping function. It first compares $E_\theta(s)$ to the input state s through a reconstruction-based discrepancy, treating well-reconstructed states as more expert-consistent and poorly reconstructed states as less consistent. It then maps this discrepancy into reward through thresholding and shaping. Concretely,

$$g(E_\theta(s), s) = \kappa \cdot \mathrm{clip}\left(f\left(\frac{\|E_\theta(s) - s\|^2 - L_{\min}}{L_{\max} - L_{\min}} \right), 0, 1 \right), \tag{2}$$

where $L_{\min}$ and $L_{\max}$ define the reward-shaping range, f is a monotone shaping function, and κ is a scaling factor. In our experiments, we use

$$f(x) = e^{-qx}, \tag{3}$$

where q controls how sharply reward decays as reconstruction error increases.

This decomposition is useful because the learned expert-similarity model and the deployed reward are distinct objects. Once E_θ has been learned, the shaping behavior can be modified through g without relearning the underlying expert-similarity model. In this sense, E_θ determines *what* structure is captured from weak expert data, while g determines *how* that structure is turned into a reward.

B. Mixture of autoencoder experts

In our experiments, we instantiate E_θ as a mixture of autoencoder experts because locomotion often has distinct behavioral phases (e.g., getting into a stable position and moving forward as fast as possible) that a single autoencoder often fails to fully capture. Given a state s, each expert k produces a reconstruction $\hat{s}_k$, and a gating network produces mixture weights $\pi_k(s)$ satisfying $\sum_{k=1}^{K} \pi_k(s) = 1$. The expert-model output is then the aggregated reconstruction

$$E_\theta(s) = \sum_{k=1}^{K} \pi_k(s) \hat{s}_k. \tag{4}$$

The role of this model is to provide an expert-conditioned description of the input state. States that are more consistent with the expert data should be reconstructed more faithfully by the mixture, but the expert-similarity model itself does not specify how reconstruction quality is translated into reward; that is the role of g.

C. Successor-state shaping for exploration

We use the learned signal as an intrinsic reward applied to successor states. After taking action a in state s and reaching successor state s', the agent receives the combined reward

$$r(s, a, s', t) = r_{\mathrm{env}}(s, a, s') + \beta(t)\, g(E_\theta(s'), s'). \tag{5}$$

The coefficient $\beta(t)$ controls how strongly the learned signal influences behavior over training. Early in learning, larger values bias the agent toward expert-supported regions, improving guidance when learning from scratch is inefficient or risky. Later, $\beta(t)$ is reduced, allowing the agent to rely more on the environment reward and improve beyond imperfect expert data. This gives an explicit handle on the exploration–exploitation trade-off.

Applying the signal to s' rather than s is deliberate. A current-state reward $g(E_\theta(s), s)$ is identical across all actions available from the same state at the level of the immediate reward and therefore provides weaker local directional guidance. By contrast, $g(E_\theta(s'), s')$ directly rewards transitions into expert-supported regions. A state only reward function can also be implemented via Potential-based Reward shaping (PBRS) [28], [29] by defining a potential

$$\Phi(s) = g(E_\theta(s), s), \tag{6}$$

then a PBRS formulation would use

$$r_{\mathrm{PBRS}}(s, a, s', t) = r_{\mathrm{env}}(s, a, s') + \beta(t)\left(\gamma \Phi(s') - \Phi(s)\right). \tag{7}$$

Under the PBRS assumptions, this biases learning while preserving the optimal policy induced by r_{env}, but for our case, we want a stronger bias towards expert-supported regions.

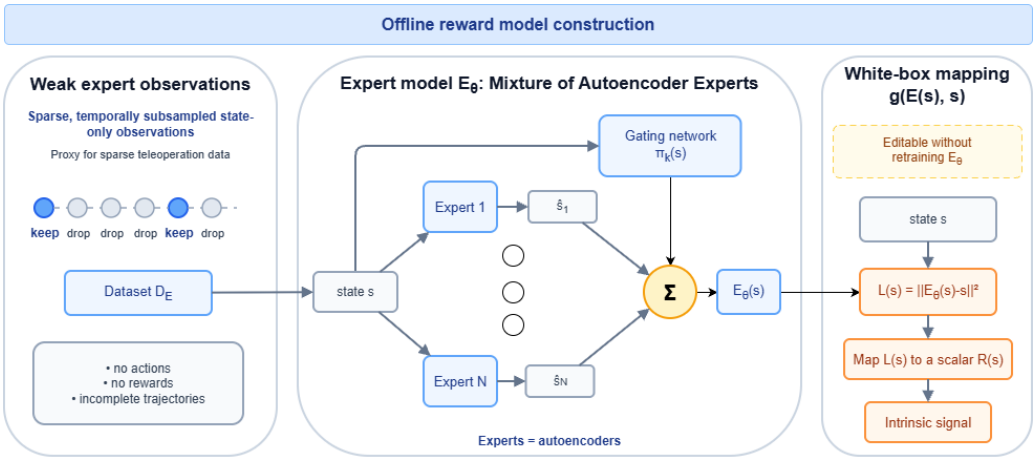

Fig. 1: Overview of MOE-GUIDE. Sparse, temporally subsampled state-only expert observations are used offline to learn an expert-similarity model E_θ, instantiated here as a mixture of autoencoder experts. At deployment, a separate white-box mapping function $g(E_\theta(s), s)$ converts reconstruction-based expert consistency into an intrinsic reward, which is applied to successor states to bias exploration toward expert-supported regions.

IV. Experiments

We evaluate MoE-GUIDE with Soft Actor-Critic (SAC) [30] on five MuJoCo continuous-control benchmarks: Swimmer, Hopper, Walker2d, HalfCheetah, and Ant. Each environment is paired with a limited set of expert demonstrations: one for Swimmer, four for Hopper, and ten for Walker2d, HalfCheetah, and Ant. To place the method in the weak-data regime, demonstrations are sparsified by recording only every fifth state, yielding sparse, state-only observations with incomplete coverage. In some cases, MoE-GUIDE uses "pretraining" for a limited number of episodes, starting in a sampled state from the dataset. Results are averaged over five random seeds, and shaded regions denote standard deviation. We evaluate MoE-GUIDE in three settings: strong expert demonstrations in intrinsic-only and dense rewards, and imperfect expert experiments additionally in sparse rewards. Across experiments, the autoencoder bottleneck was set slightly below half the state dimension and the number of experts kept below five. The mapping used an exponential form with steepness $q = 100$, except in intrinsic-only runs where $q = 200$ imposed a stricter expert-support prior; $L_{\min}$ and $L_{\max}$ depend on the learned expert model and κ controls the environment-dependent strength of exploration bias. We used random policy rollouts to find the usable range of $L_{\min}$ and $L_{\max}$ and an appropriate steepness, then swept three $L_{\min}$ values in the intrinsic-only setting to find a signal that captures the expert well without overrewarding OOD data. In the dense setting, we tested two to three scaling factors; these final settings were transferred unchanged to the sparse-reward setting.

a) Baseline selection.: Since this paper presents preliminary results in the weak-data exploration regime, we focus on baselines that operate under the same information and interaction budget: sparse, state-only expert observations without action labels, full trajectories, auxiliary transition datasets, or additional reward-construction interaction. Other demonstration-based methods require richer trajectories, transitions, occupancy estimates, or reward-construction interaction, and this paper does not claim to beat them under their assumptions. Our main comparisons are (1) extrinsic reward only (ER-only), (2) extrinsic reward with pretraining on demonstration states (ER+pretraining), (3) intrinsic reward from the learned reward model with pretraining (IR+pretraining), and (4) the full method combining extrinsic and intrinsic reward (MoE-GUIDE). For context, we also report generic intrinsic-motivation baselines such as autoencoder-based, RND and ICM following the guidelines from [31]; unlike MoE-GUIDE, these methods do not use weak expert data.

Strong expert. The results of leveraging strong expert data are shown in Figure 2. In Swimmer, Walker2d, and Ant, MoE-GUIDE reliably improves over ER-only. In Hopper, both IR+pretraining and MoE-GUIDE reach expert-level performance during training, indicating that the learned intrinsic reward provides particularly effective guidance in this environment. By contrast, in HalfCheetah, the extrinsic reward already guides exploration effectively toward high-return regions, leaving less room for improvement from additional intrinsic shaping. As we can see, standard non-data-leveraging intrinsic motivation methods, do not yield much benefit in a dense rewards setting, even hindering progress in most cases. While MoE-GUIDE keeps providing useful guidance even if the environment reward is dense.

Imperfect expert. As shown in Figure 3, MoE-GUIDE consistently improves over the imperfect expert in all environments. For HalfCheetah, we intentionally use a very poor-performing expert to test whether even low-quality weak guidance can remain useful. In this case, MoE-GUIDE does not exceed ER-only, but it achieves comparable and more stable learning.

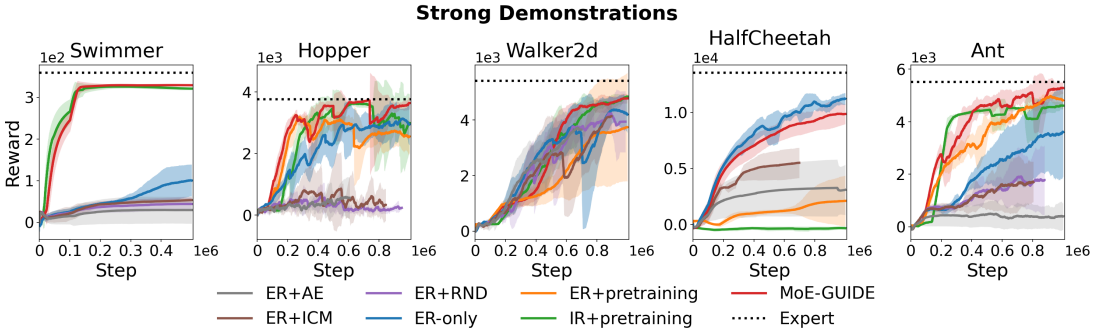

Fig. 2: Strong-expert setting on five MuJoCo locomotion environments. MOE-GUIDE is compared with ER-only, ER+pretraining, and IR+pretraining. Mean over five seeds; shaded regions show standard deviation.

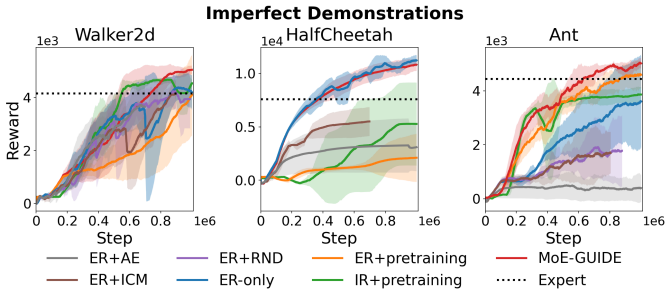

Fig. 3: Imperfect-expert setting across five MuJoCo environments. MOE-GUIDE is compared against ER-only, ER+pretraining, and IR+pretraining.

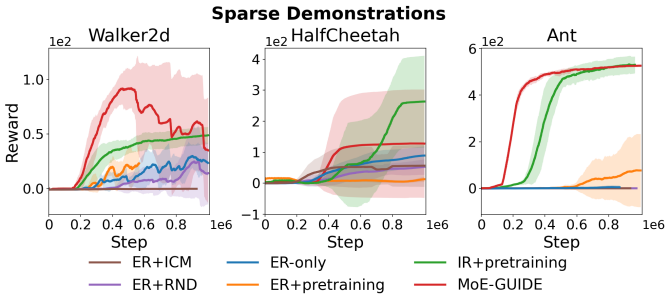

Fig. 4: Sparse-reward, partially observable setting. The same learned reward model is reused unchanged from the dense-reward case.

Sparse rewards. Figure 4 presents results in modified MuJoCo environments where reward is only provided at sparse checkpoints, making exploration substantially harder. The agent's current position (x-coordinate) is omitted from both observations and demonstrations, resulting in partial observability. Importantly, the dense- and sparse-reward settings use the exact same learned reward model: the expert-similarity model E and mapping function g are reused unchanged, without retraining or redesigning the reward artifact for the sparse setting. Despite never crossing checkpoints during pretraining, MOE-GUIDE substantially outperforms the main baselines in these sparse, partially observable environments. For HalfCheetah, however, extrinsic rewards help early in training, while agents optimizing only intrinsic rewards eventually outperform the combined objective. This reuse is important because it tests whether the same weak-data-trained artifact transfers unchanged across reward regimes.

b) Ant ablations.: Figure 5 summarizes the main Ant ablations. First, MOE-GUIDE remains informative under substantial weak-data degradation: as demonstrations become sparser, performance declines only gradually and remains well above ER-only. Second, the decay-rate study shows that Ant prefers a slow decay of expert influence, indicating that sustained expert guidance remains useful in this environment. Third, the mapping-threshold study shows that overly permissive thresholds degrade extrinsic performance, whereas tighter thresholds yield much stronger returns. Taken together, these ablations show that weak expert observations can remain useful even under severe sparsification, but that the practical quality of the learned artifact depends strongly on reward shaping and schedule selection.

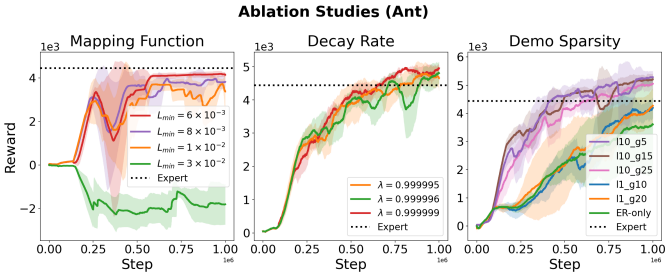

Fig. 5: Ant ablations. MOE-GUIDE remains useful under severe demonstration sparsity, but performance depends on decay rate and reward-mapping thresholds.

c) Sweep analysis.: Figure 6 summarizes a sweep over 349,313 checkpoints from 9,434 expert-model configurations, varying bottleneck dimension, number of experts, entropy regularization, and random seed. Because downstream RL evaluation at this scale is infeasible, we analyze the learned artifacts using separation metrics. *Lowest separation* reports the worst separation against 12 held-out behavior datasets generated by agents with substantially different behaviors, making

it a conservative selectivity measure. Training loss is not a complete surrogate for artifact quality: high-loss models show highly variable separation, while poor artifacts become much rarer in the low-loss regime. Capacity parameters, such as the number of experts and the bottleneck dimension, yield only modest median gains with substantial spread. Random-OOD separation, computed against random-policy data, preserves the main qualitative trends and provides a practical proxy when curated non-expert trajectories are unavailable.

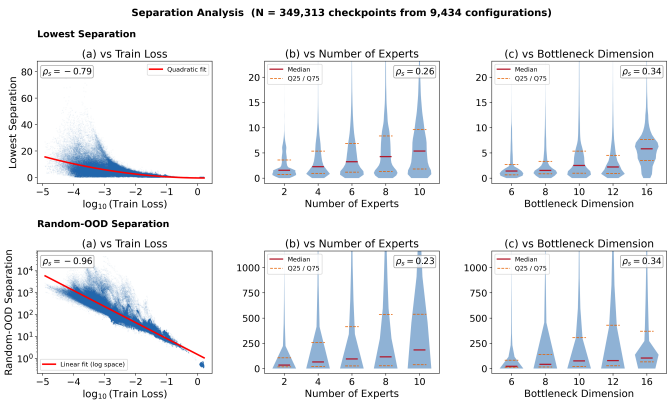

Fig. 6: Sweep analysis over expert-model configurations. *Lowest separation* corresponds to the worst separation on a held-out set of 12 test sets. Lower train loss correlates with a more reliable regime, while random-policy OOD separation preserves the main qualitative trends.

Discussion. Overall, the results show that MoE-GUIDE provides useful exploration guidance across different environments and expert qualities, with the strongest gains when extrinsic rewards are sparse, partial, or otherwise insufficient for efficient exploration. In dense-reward environments such as HalfCheetah, extrinsic reward already provides strong guidance, so the advantage of additional support-based shaping is smaller. We also observe that pretraining on demonstration data can be useful, but without continued guidance afterward, it may hinder downstream learning.

A practically important result is the reusability of the learned artifact. The exact same learned expert-similarity model E and mapping function g, apart from an exponential steepness change in the intrinsic-only setting, are deployed unchanged in intrinsic-only training, standard dense-reward training, and sparse-reward training, without retraining or redesigning the reward artifact for each case. This suggests that weak-data-trained evaluators can function as reusable exploration priors rather than one-off reward constructions tied to a single reward regime. Taken together, these results support the view that weak expert observations can serve as a practical and reusable exploration artifact even when they are too sparse and incomplete for more standard uses of demonstrations.

V. Discussion and Conclusion

We presented MOE-GUIDE, a method for exploration from unlabeled, incomplete, and imperfect expert observations. Rather than using expert data as direct policy supervision or as the basis for full reward recovery, we learn an expert-similarity model from static observations and deploy it through a separate mapping function that converts expert consistency into a successor-state shaping reward. This yields an exploration prior based on weak expert evidence: less ambitious than recovering a task-complete reward or explicit constraint set, but also requiring weaker assumptions, since it does not depend on action labels, complete trajectories, explicit correspondence, or joint reward–policy optimization during artifact construction.

A central design choice is the separation between the expert-similarity model $E(s)$ and the mapping function $g(E(s), s)$. The expert-similarity model captures structure from weak observations, while the mapping function determines how that structure is translated into reward. In our current instantiation, this mapping is white-box and reconstruction-based, making the shaping behavior explicit and editable after the expert-similarity model has been learned. More broadly, this separation exposes a design space in which expert-similarity modeling and reward deployment can be modified separately, enabling post hoc calibration, patching, and refinement of the learned artifact.

Across state-based locomotion benchmarks, the results show that weak expert observations can provide useful guidance for exploration under substantially weaker assumptions than most demonstration-based policy or reward-learning methods require. The strongest gains appear when extrinsic reward is sparse, partial, or otherwise insufficient for efficient exploration, while dense-reward settings such as HalfCheetah leave less room for additional support-based shaping. At the same time, the experiments delimit the scope of the paper. The learned signal is a support-based bias toward regions represented in the expert data, and its usefulness depends on the informativeness and alignment of that data. If the observations are too sparse, weakly aligned, or systematically misleading, the signal may provide poor guidance or reinforce suboptimal regions. The present evidence is also limited to state-based continuous-control benchmarks, so the broader significance is not that we solve weak observational learning in general, but that we begin to make a class of weak expert data operational for exploration.

Future work should exploit the separation between E and g more directly. One direction is to improve pre-deployment patching and calibration of the learned artifact [10], [32]. Another is to replace fixed reward-shaping ranges with adaptive or dynamic calibration of $L_{\min}$ and $L_{\max}$, so that the mapping function can adjust to the learned expert model and the distribution of encountered states rather than relying on static thresholds. It would also be valuable to extend the framework to richer and less reliable observation modalities, including image-based, partially observed, and degraded inputs, and to study more local forms of support-aware filtering or safer exploration in settings where weak expert data is available but direct supervision remains impractical.

ACKNOWLEDGMENT

This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" program.

REFERENCES

[1] R. S. Sutton, A. G. Barto, *et al.*, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[2] R. McCarthy, D. C. Tan, D. Schmidt, F. Acero, N. Herr, Y. Du, T. G. Thuruthel, and Z. Li, "Towards generalist robot learning from internet video: A survey," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[3] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[4] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6292–6299, IEEE, 2018.

[5] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[6] T. L. Paine, C. Gulcehre, B. Shahriari, M. Denil, M. Hoffman, H. Soyer, R. Tanburn, S. Kapturowski, N. Rabinowitz, D. Williams, *et al.*, "Making efficient use of demonstrations to solve hard exploration problems," *arXiv preprint arXiv:1909.01387*, 2019.

[7] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, (New York, NY, USA), p. 1, Association for Computing Machinery, 2004.

[8] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[9] S. Deshpande, R. Walambe, K. Kotecha, G. Selvachandran, and A. Abraham, "Advances and applications in inverse reinforcement learning: a comprehensive review," *Neural Computing and Applications*, pp. 1–53, 2025.

[10] E. Malomgré and P. Simoens, "Interactionless inverse reinforcement learning: A data-centric framework for durable alignment," *arXiv preprint arXiv:2602.14844*, 2026.

[11] R. Burnwal, H. Mehta, N. P. Bhatt, and B. Ravindran, "Learning from observation: A survey of recent advances," *arXiv preprint arXiv:2509.19379*, 2025.

[12] A. Aubret, L. Matignon, and S. Hassas, "An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey," *Entropy*, vol. 25, no. 2, p. 327, 2023.

[13] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," *Advances in neural information processing systems*, vol. 29, 2016.

[14] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International conference on machine learning*, pp. 2778–2787, PMLR, 2017.

[15] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," *arXiv preprint arXiv:1810.12894*, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[18] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078–10093, 2022.

[19] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*, pp. 709–727, Springer, 2022.

[20] A. Camacho, I. Gur, M. L. Moczulski, O. Nachum, and A. Faust, "Sparsedice: Imitation learning for temporally sparse data via regularization," in *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021.

[21] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.

[22] H. Wei, C. Chen, C. Liu, G. Zheng, and Z. Li, "Learning to simulate on sparse trajectory data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 530–545, Springer, 2020.

[23] M. Sun and X. Ma, "Adversarial imitation learning from incomplete demonstrations," *arXiv preprint arXiv:1905.12310*, 2019.

[24] D. Xu, F. Zhu, Q. Liu, and P. Zhao, "Arail: Learning to rank from incomplete demonstrations," *Information Sciences*, vol. 565, pp. 422–437, 2021.

[25] T. Ni, H. Sikchi, Y. Wang, T. Gupta, L. Lee, and B. Eysenbach, "f-irl: Inverse reinforcement learning via state marginal matching," in *Conference on Robot Learning*, pp. 529–551, PMLR, 2021.

[26] G. Chaudhary and L. Behera, "From novelty to imitation: Self-distilled rewards for offline reinforcement learning," *arXiv preprint arXiv:2507.12815*, 2025.

[27] Y. Ma, A. Shen, D. Jayaraman, and O. Bastani, "Versatile offline imitation from observations and examples via regularized state-occupancy matching," in *International Conference on Machine Learning*, pp. 14639–14663, PMLR, 2022.

[28] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99, pp. 278–287, Citeseer, 1999.

[29] E. Wiewiora, "Potential-based shaping and q-value initialization are equivalent," *Journal of Artificial Intelligence Research*, vol. 19, pp. 205–208, 2003.

[30] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, Pmlr, 2018.

[31] M. Yuan, R. C. Castanyer, B. Li, X. Jin, W. Zeng, and G. Berseth, "Rl-explore: Accelerating research in intrinsically-motivated reinforcement learning," *arXiv preprint arXiv:2405.19548*, 2024.

[32] E. Malomgré and P. Simoens, "The alignment flywheel: A governance-centric hybrid mas for architecture-agnostic safety," *arXiv preprint arXiv:2603.02259*, 2026.