AEGIS: ADVERSARIAL EXPLORATION FOR GENERALIZED IMPROVED STATE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-free pretraining (RFPT) seeks transferable state representations and exploration behaviors before task-specific rewards are available, but existing approaches treat exploration and representation learning as independent processes. AEGIS combines them as an approximate bilevel Stackelberg problem between (i) a latent dynamics model that minimizes the prediction error and (ii) an exploration policy that uncovers states that expose its weaknesses. The policy is driven by an intrinsic reward combining intra-episode diversity with novelty toward states the dynamics model has not yet fitted. We prove that each component of the representation model and the intrinsic reward exclude a degenerate solution the others allow, so only their combination is non-degenerate. AEGIS outperforms state-of-the-art methods with faster convergence and less redundant exploration on MiniGrid and Procgen.

1 INTRODUCTION

The success of reinforcement learning (RL) often hinges on carefully designed reward signals and extensive online interaction, both impractical in real-world settings (Sutton & Barto, 2018; Ladosz et al., 2022). Reward-free pretraining (RFPT) mitigates these challenges by decoupling representation learning from the task reward: the agent interacts with the environment without access to rewards and learns general-purpose representations, dynamics models, or exploratory policies that can adapt rapidly once a reward signal appears. This reduces the sample complexity of task learning and lets the pretrained model transfer across tasks, domains, and environments. As an online method, RFPT must balance exploration against exploitation, since incomplete or repetitive exploration yields sparse or biased representations, degrading robustness and generalization (Xie et al., 2022), and every environment interaction has a cost. Representation learning and exploration are moreover interdependent: uninformative transition data yield biased latent embeddings, and policies built on biased embeddings cannot steer exploration toward novel, task-relevant states (Badia et al., 2020; Botteghi et al., 2022; Amodei et al., 2016), trapping the agent in already-explored regions. Addressing this self-reinforcing cycle requires an exploration policy that covers the critical regions of the state space while learning accurate representations and diverse behaviors without extrinsic reward.

To mitigate these issues, we introduce *AEGIS* (Adversarial Exploration for Generalized Improved State Representations), an RL framework that refines representations through adversarially guided exploration. *AEGIS* employs two complementary intrinsic reward signals: a global component that drives the agent toward states that expose weaknesses in the current representation model, and a local component that encourages diverse behaviors within each episode. Together, they promote non-degenerate latent representations and richer exploratory skills, accelerating policy learning as illustrated in Figure 1. Intrinsic motivation continues to benefit learning even after extrinsic rewards become available, further refining the final policy. Overall, this paper's contributions are:
1) Frame RFPT as an approximate bilevel Stackelberg optimization problem.
2) Local and global intrinsic reward components that promote efficient exploration and convergence.
3) A novel algorithm for learning state representations and diverse behaviors through adversarial exploration that is compatible with both online RFPT and policy learning.
4) Show that each component of the representation model and the intrinsic reward optimizes a proxy objective with a degenerate solution in isolation, and that each pairing excludes exactly the solution its partner allows, justifying the composite design.
5) Empirical results across MiniGrid and Procgen showing superior performance against two state-of-the-art methods and four more baselines, with ablations matching the predicted failure modes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

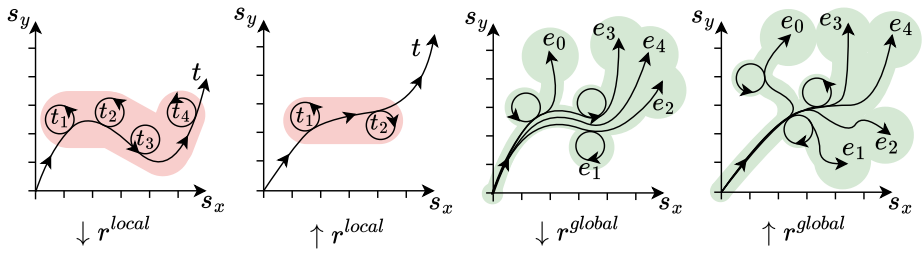


Figure 1: Intrinsic motivation of *AEGIS*. r^{local} motivates diverse behavior in an episode (e_i) while r^{global} directs exploration toward states the dynamics model has not yet fitted. Together, the two components yield diverse skills, robust representations, and less redundant exploration.

2 Background

The Markov Decision Process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ (Sutton & Barto, 2018), where $\mathcal{P}(s'|s,a)$ specifies the Markovian transition dynamics and a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maximizes the expected discounted return $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$. When the agent lacks direct access to the true state, the framework generalizes to POMDPs, where the agent receives observations $o_t \in \mathcal{O}$ from an emission distribution $q(o|s_t)$ that may alias distinct states (Kaelbling et al., 1998), making optimal behavior generally intractable (Spaan, 2012). A more tractable subset is the Block MDP (BMDP), which assumes a latent space $\mathcal{S}$ with Markovian dynamics behind a distinct, typically high-dimensional observation space $\mathcal{O}$ (Du et al., 2019). Its defining assumption is *block structure*: the emission distributions of distinct latent states have disjoint supports, $\mathrm{supp}(q(\cdot|s)) \cap \mathrm{supp}(q(\cdot|s')) = \emptyset$ for $s \neq s'$, so every observation is generated by exactly one latent state and the state is uniquely identifiable from it, decoupling the perception problem from the control problem (Stooke et al., 2021). Following Kaelbling et al. and Du et al., a POMDP reduces to a BMDP under mild conditions on the interaction history and the emission supports; raw single-step observations often violate block structure under partial observability, but the framework can be recovered by considering *sufficient observations*, i.e. histories or recurrent encodings that are sufficient statistics for the latent state.

Under these conditions, representation learning amounts to learning an encoder f_θ mapping sufficient observations into a latent space that preserves the block property, so that the POMDP is effectively unrolled into a standard MDP. Throughout, o_t denotes a sufficient observation and z_t its embedding; Section 3 splits z_t into the per-observation embedding $z_t^{obs} \in \mathcal{Z}^{obs}$ and the history-dependent embedding $z_t^{traj} \in \mathcal{Z}^{traj}$, distinguished from that point on. We work in the episodic setting with horizon H. An episode yields a trajectory $\tau = (o_0, a_0, \ldots, a_{H-1}, o_H)$, and $\mathcal{S}(\tau)$ denotes the set of distinct latent states it visits. In the environments we evaluate, transitions are deterministic, and emissions are point masses where each latent state yields a single observation; Section A uses this.

3 Methodology

AEGIS is comprised of four parts: i) the representation learning model, ii) the intrinsic reward model, iii) the RL agent, and iv) the adversarial exploration Objective.

The Representation Learning Model. To learn meaningful state representations, the encoder needs to preserve the environment's dynamics and provide useful features for the RL task. To do that, we use a Latent MDP model (LMDP) that learns a Forward Dynamics Model (FDM), an Inverse Dynamics Model (IDM), and a Discriminative Contrastive Learning Model (DSCM). This way, the representations learned preserve the homomorphic properties of BMDPs and POMDPs and, at the same time, encourage more informative embedding features (Figure 2 middle).

The encoder f_θ produces two embeddings of an observation o_t: the observation embedding z_t^{obs}, the output of its CNN block, and the trajectory-aware embedding z_t^{traj}, the output of its RNN block. The LMDP heads operate on z^{traj}. Given two consecutive observations o_t, o_{t+1} and the action a_t, the FDM h_δ takes z_t^{traj}, a_t and predicts the next embedding $\hat{z}_{t+1}^{traj}$, scored against the true z_{t+1}^{traj} using Mean Squared Error (MSE, Eq. (2)). The IDM q_ζ predicts the action $\hat{a}_t$ that caused the

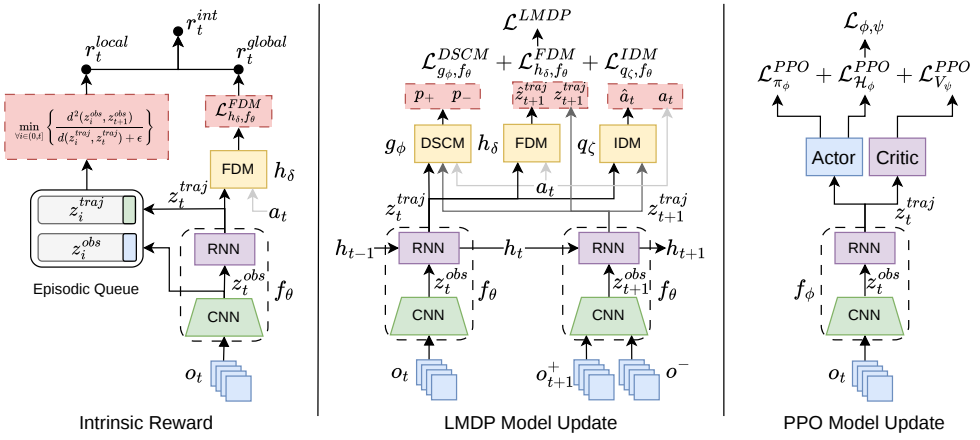


Figure 2: Overview of Aegis. The computation of the intrinsic rewards for adversarial exploration and representation learning (left), the LMDP modules (middle), and the RL modules used by PPO (right).

transition, scored against a_t using Categorical Cross-Entropy (CCE, Eq. (3)). The DSCM g_ω takes $z_t^{traj}, z_{t+1}^{traj}, a_t$ and predicts whether z_{t+1}^{traj} comes from the true next observation (o_{t+1}^+) or from a fake one (o^-), scored using Binary Cross-Entropy (BCE, Eq. (4)). The positive and negative terms of Eq. (4) are averaged with equal weight, so the effective class prior is $1/2$ regardless of how many negatives are discarded as duplicates of the positive. For the fake observations, we sample from a queue of recent novel past experiences, novel with respect to the running mean of r^{int}, as in (Wan et al., 2023). We calculate the total loss for the LMDP model as:

$$\mathcal{L}^{LMDP} = w_F \cdot \mathcal{L}^{FDM}_{h_\delta, f_\theta} + w_I \cdot \mathcal{L}^{IDM}_{q_\zeta, f_\theta} + w_D \cdot \mathcal{L}^{DSCM}_{g_\omega, f_\theta} \tag{1}$$

$$\text{where:} \quad \mathcal{L}^{FDM}_{h_\delta, f_\theta} = \mathcal{J}^{MSE}(h_\delta(z_t^{traj}, a_t),\, z_{t+1}^{traj}), \tag{2}$$

$$\text{where:} \quad \mathcal{L}^{IDM}_{q_\zeta, f_\theta} = \mathcal{J}^{CCE}(q_\zeta(z_t^{traj}, z_{t+1}^{traj}),\, a_t), \tag{3}$$

$$\text{and:} \quad \mathcal{L}^{DSCM}_{g_\omega, f_\theta} = \mathcal{J}^{BCE}(g_\omega(z_t^{traj}, z_{t+1}^{traj}, a_t),\, p^{+/-}) \tag{4}$$

where $z_t^{obs} = f_\theta^{obs}(o_t)$ is the CNN output and $z_t^{traj} = f_\theta^{traj}(o_{0:t})$ the RNN output, with $\mathcal{Z}^{traj} = \mathbb{R}^d$. We use $(w_F, w_I, w_D) = (0.5, 0.5, 1)$ in all experiments. We write $\theta = (\theta_{enc}, \delta, \zeta, \omega)$ for the LMDP parameters collectively and f_θ for its encoder component, so that $\mathcal{L}^{LMDP}$ and each of its terms are functions of θ. The LMDP maintains its own CNN and RNN modules, architecturally identical to but separate from those of the policy; each is updated only by its own objective.

The Intrinsic Reward Model. In the pretraining task where the extrinsic reward is not available, the goal is to explore the state space and find states that are unfamiliar to the LMDP model. For this reason, we use two intrinsic reward terms: a local reward and a global reward (Figure 2 left). The local rewards are computed episodically using observations collected within the same episode, while global rewards direct the agent toward transitions the dynamics model has not yet fitted. We calculate the total intrinsic reward by multiplying the two reward components:

$$r_t^{int} = \alpha \cdot r_t^{local} \cdot r_t^{global} \tag{5}$$

where α is the intrinsic reward impact coefficient. Depending on the true reward scale, α is adjusted so r_t^{int} is two to three levels of magnitude smaller than the extrinsic reward. Multiplying the two reward components is inspired by amplitude modulation from signal processing. We test each component's contribution in our ablation study presented in Section 5.

The local objective is $J = d(o_i, o_{t+1}) \cdot \mathbb{I}\big(d(o_i, o_{t+1});\, a_t \mid s_t, s_i\big)$, the distance between two observations weighted by the conditional mutual information between that distance and the action taken. Wan et al. (2023) bound it below by a tractable surrogate, under deterministic transitions and emissions and the assumption that the observation distance is exponentially distributed with mean $d(s_i, s_t)$:

$$J \;\geq\; d(o_i, o_{t+1}) \cdot \mathbb{E}_{s_t, s_i, a_t}\left(\log\big(d(s_i, s_t)\big) + \frac{d(o_i, o_{t+1})}{d(s_i, s_t)}\right) \tag{6}$$

Two simplifications make this usable as a per-step reward: we keep only the dominating term $d(o_i, o_{t+1})/d(s_i, s_t)$, and replace the expectation by a minimum over the episodic history. Neither step preserves the inequality, so Eq. (7) is an approximation of Eq. (6) rather than a bound on J. Since s_t is not observable, z^{traj} serves as its proxy:

$$r_t^{local} = \min_{\forall i \in [0,t)} \left\{ \frac{d^2(z_i^{obs}, z_{t+1}^{obs})}{d(z_i^{traj}, z_t^{traj}) + \epsilon} \right\}, \qquad r_0^{local} = 0 \tag{7}$$

where $d(\cdot, \cdot)$ is the Euclidean distance, applied to observations and states in Eq. (6) and to their embeddings in Eq. (7), ϵ is a small constant for numerical stability, and the minimum over an empty history at $t = 0$ is taken to be zero. The two distances in Eq. (7) play different roles. The denominator is a distance in $\mathcal{Z}^{traj}$ standing in for the distance between the underlying states, so it is only meaningful if that geometry is constrained; the numerator is a distance in $\mathcal{Z}^{obs}$ and requires only that distinct observations receive distinct embeddings. Theorem 1 in Section A shows that the LMDP objective of Eq. (1) supplies both. Eq. (7) is the episodic reward of Wan et al. (2023); our contribution at the reward level is r^{global} and the multiplicative composition, and at the representation level, the objective that constrains the latent geometry this reward consumes.

We define the global reward r^{global} as the novelty difference of the current step and the previous. A transition is novel when it causes the FDM model to have errors, meaning that the corresponding states fall outside the predictive abilities of the FDM model.

$$v_t = \left\| h_\delta(z_t^{traj}, a_t) - z_{t+1}^{traj} \right\|^2 \tag{8}$$

$$r_t^{global} = \max(v_t - 0.5 \cdot v_{t-1}, 0) + 1 \tag{9}$$

where the $+1$ term ensures $r^{global} \geq 1$. $\mathcal{L}^{FDM}_{h_\delta, f_\theta} = \mathbb{E}[v_t]$ and v_t is its per-transition value. We use novelty differences instead of simple novelty scores in Eq. (9) for more breadth-first-like global exploration as displayed by (Zhang et al., 2021).

Representation Learning through Adversarial Exploration. *AEGIS* trains the LMDP model on data generated by a policy that is executed and updated using the on-policy PPO algorithm (Schulman et al., 2017). The intrinsic reward (Eq. (5)) is optimized by the policy and encourages the agent to develop strategies that systematically explore the state space and uncover informative data that are out-of-distribution for the LMDP model. The collected data are fed back to the LMDP model for representation learning. Algorithm 1 in App. D describes the full algorithmic logic in greater detail.

The LMDP and the policy maintain separate encoders, f_θ and f_ϕ, architecturally identical but updated only by their own objectives, so the two interact only through the intrinsic reward and the data the policy collects. The policy acts on $f_\phi(o_t)$, while both intrinsic reward components are computed from f_θ and the LMDP heads, so r^{int} depends only on θ. For each newly encountered state, the agent initially receives a high intrinsic reward, which decreases as the state becomes familiar within the same episode; states seen across different episodes may receive a high local reward but a low global one. To maximize both components, the agent must therefore seek transitions the dynamics model has not yet fitted while remaining geometrically distant from states already visited within the episode. Section A formalizes why neither component alone produces this behavior.

We term this strategy *Adversarial Exploration* and formulate it as an approximate bilevel Stackelberg optimization problem (Nisan et al., 2007), where the LMDP model acts as the leader defining the intrinsic reward landscape and the exploration policy acts as the follower:

$$\phi^*(\theta) = \arg\max_\phi \mathbb{E}_{\pi_\phi}\left[r^{int}(\theta)\right], \text{where } \theta^* \approx \arg\min_\theta \mathcal{L}^{LMDP}(\theta; \mathcal{D}(\pi_{\phi^*(\theta)})) \tag{10}$$

where θ denotes the LMDP parameters collectively, as defined in Subsection 3. In practice, this is approximated via alternating optimization with timescale separation: the policy is updated through multiple PPO epochs, and the LMDP model is updated more slowly on the induced data, giving an implicit best-response dynamic without explicit meta-gradient computation.

Because the two levels share no parameters, their only coupling is through r^{int} and the induced data distribution $\mathcal{D}(\pi_\phi)$: the global reward is directly coupled to the LMDP prediction error, creating a competitive interaction with the representation model (giving the framework its adversarial nature), while the local reward component is independent of the LMDP objective and promotes intra-episode

diversity. As a result, the overall interaction is best characterized as a general-sum leader–follower game rather than a strictly zero-sum minimax problem. The zero-sum alternative is also degenerate: with $r^{int} = r^{global}$ and the LMDP reduced to its FDM term, the representation can trivially minimize that error by collapsing, a solution that Lemma 1 in Section A shows the full objective excludes and that we observe empirically in the min-max ablation of Section 5.

Theoretical justification of the components. Both the representation model and the intrinsic reward of *AEGIS* are built from several components, which might appear assembled empirically rather than derived. We show in App. A that the composition follows a single structural principle: *each component optimizes a proxy objective that has a degenerate solution when used in isolation, and each pairing eliminates exactly the degenerate solution that the other component tolerates.* We instantiate this principle twice, once at the representation level and once at the reward level, characterizing which failure modes each component provably excludes; each result corresponds to an ablation reported in Section 5. At the representation level, the forward model constrains the latent geometry but permits a collapsed encoder, which the inverse model and the discriminator rule out (Lemma 1, Lemma 2, Theorem 1). At the reward level, the local reward is non-degenerate only because a revisit earns nothing (Proposition 1), while the global reward alone does not enforce the policy to seek states the model is weak on; only their product excludes both failure modes (Theorem 2).

4 EVALUATION

We evaluate *AEGIS* against other popular methods in the area in terms of policy performance, convergence speed, and exploration redundancy.

Experimental Setup. We evaluate *AEGIS* in RFPT settings by splitting the interaction budget into a pretraining and a training phase. During pretraining, the environment's extrinsic rewards are nullified, so the agent relies only on its exploration-driven policy and intrinsic rewards; *AEGIS* has no prior knowledge of the environment's true reward and, consequently, of what the task is about. We consider four allocations of the total budget: no pretraining (NPT), one quarter (1QPT), one half (HPT), and three quarters (3QPT). The main text reports NPT, which isolates the intrinsic reward acting as an exploration bonus alongside the task reward, and HPT, our default regime; the remaining allocations are reported in App. C and show the same ordering. We compare *AEGIS* against five exploration methods: RND, ICM, NGU, DEIR, and NovelD, along with the base PPO algorithm. All methods use PPO with identical optimization hyperparameters (learning rate, clipping, batch size, and number of epochs) so that observed differences are attributable to the intrinsic reward rather than the policy optimizer, following the protocol of Wan et al. (2023); PPO is also NovelD's original optimizer (Zhang et al., 2021), so this is not a choice we impose on it. We follow the original implementation of these methods where available, while for NGU we follow the PPO adaptation used by Wan et al. (2023). Method-specific intrinsic-reward parameters follow the tuning of Wan et al. for the environments we use, and the defaults proposed in the respective original publications otherwise. The hyperparameters of *AEGIS* were selected once on DoorKey-8x8 and then frozen for every other environment and benchmark, so all results outside that task are obtained under zero-shot transfer of the configuration. No tuning was performed on any environment reported in Section 4 or on the component ablations of Section 5. For fair comparisons, we report the true environment's rewards, ignoring the effects of intrinsic motivation for each method. All curves show the median over seeds with the interquartile range shaded: ten seeds for Minigrid and three for Procgen.

Experiments in Minigrid. Minigrid is a procedurally-generated RL benchmark of sparse-reward grid-world games requiring complex, ordered use of objects to reach the goal (Chevalier-Boisvert et al., 2023), which makes it well-suited to testing exploration. We tested *AEGIS* on three challenging environments: **BlockedUnlockPickup**, **DoorKey-16x16**, **KeyCorridor-S6R3**, described in App. C. Figure 3 reports agent performance in terms of true environment rewards, with the vertical dotted line marking the phase boundary under HPT. The two rows isolate what the pretraining contributes. Under NPT, *AEGIS*, DEIR and NovelD are closely matched on DoorKey-16x16 and KeyCorridor-S6R3, with *AEGIS* ahead only on BlockedUnlockPickup. Under HPT, the picture separates: on DoorKey-16x16 *AEGIS* reaches near-optimal return several hundred thousand steps before DEIR and NovelD. This is driven entirely by the intrinsic reward, showing *AEGIS* can perform the complex action sequences needed to reach remote subspaces without ever observing the task reward. BlockedUnlockPickup is the sharpest case: *AEGIS* reaches a return above 0.9, whereas DEIR degrades from 0.87 under NPT to near zero under HPT, and ICM remains below 0.4 throughout.

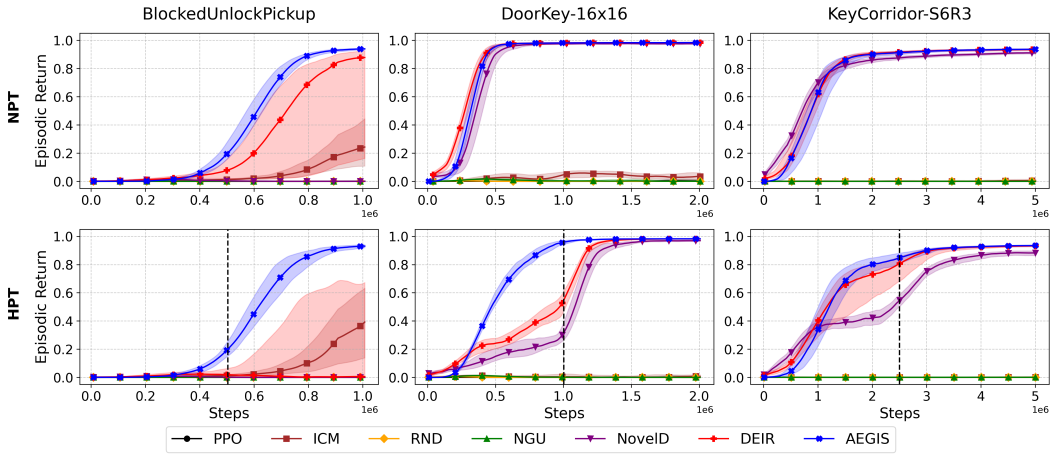


Figure 3: Episodic return in Minigrid. Median and interquartile range over ten seeds. The vertical dotted lines separate the pretraining phase from the training phase.

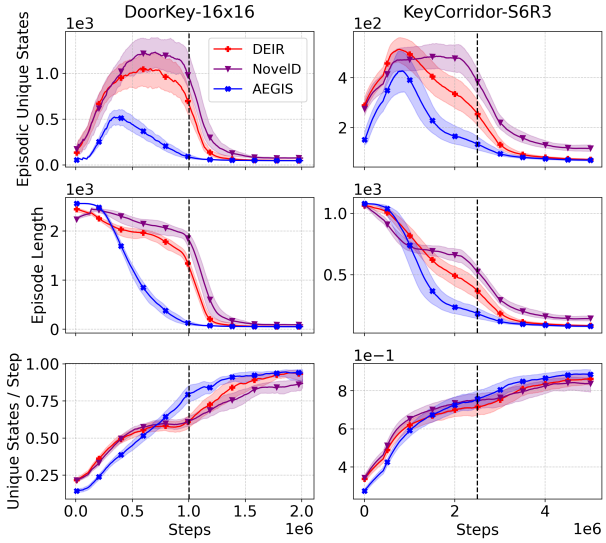


Figure 4: Episodic unique states, episode length, and unique states per step on **DoorKey-16x16** and **KeyCorridor-S6R3**. Median and IQR over 10 seeds.

	DoorKey 16x16	KeyCorridor S6R3
T_{90}	-27.7	-15.0
$\|S\|_t$	-86.0	-47.0
$\|\tau\|_t$	-89.8	-50.1
$\|S\|_t/\|\tau\|_t$	+26.7	+5.5

Table 1: Percentage difference (%) between *AEGIS* and second best method (DEIR). T_{90} is the number of steps to reach 90% of the best final return achieved by any method; the remaining rows are measured at the pretraining–training split, averaged over a window of $\pm 5\%$ of total steps: distinct states per episode $|S|_t$, episode length $|\tau|_t$, and distinct states per step $|S|_t/|\tau|_t$. Negative values favor *AEGIS* on the first three rows.

Since *AEGIS*, DEIR, and NovelD all converge to near-optimal return on DoorKey-16x16 and KeyCorridor-S6R3 and the final score does not distinguish them, we measure how quickly each reaches it, reporting T_{90}, the number of steps needed to attain 90% of the best final return achieved by any method (Table 1). On DoorKey-16x16, *AEGIS* reaches the threshold after 8.4×10^5 steps against 1.16×10^6 for DEIR and 1.27×10^6 for NovelD, a 27.7% reduction over the closest baseline, with the per-seed crossings of *AEGIS* and DEIR falling in non-overlapping interquartile ranges. Notably, this crossing occurs before the phase boundary, so *AEGIS* attains near-optimal behavior during the reward-free phase alone. On KeyCorridor-S6R3 the margin is 15.0% over DEIR and the crossing falls at the phase boundary rather than before it; NovelD fails to reach the threshold on one seed there.

AEGIS appears to explore substantially less by one measure: at the phase boundary it visits 90 distinct states per episode on DoorKey-16x16 against 641 for DEIR, an 86.0% reduction, and 133 against 252 on KeyCorridor-S6R3 (Figure 4, top row). This count is ambiguous on its own, since an agent that reaches the terminating state sooner touches fewer states regardless of how it explores. The middle row resolves the ambiguity: episode length falls by a comparable proportion, 89.8% and 50.1%, so most of this reduction is accounted for by shorter episodes rather than by narrower exploration. Normalized for episode length, *AEGIS* visits *more* distinct states per step than DEIR:

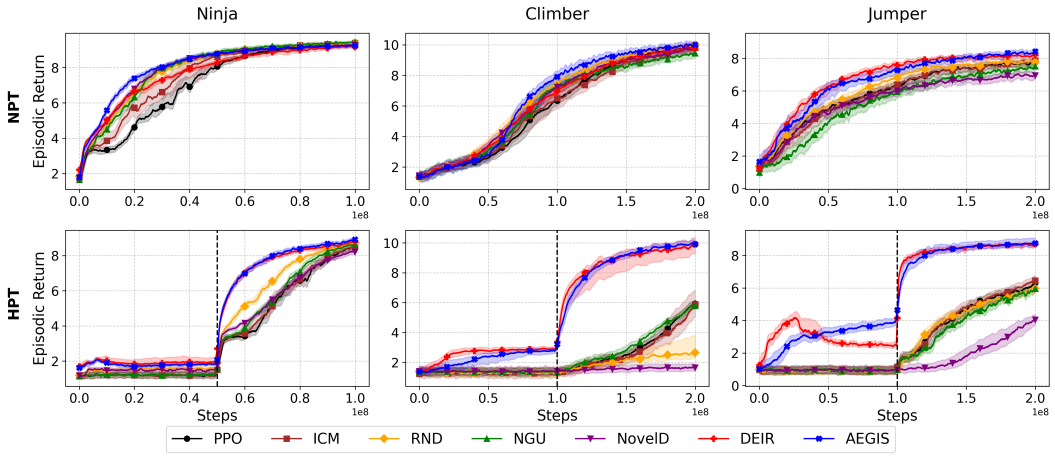


Figure 5: Episodic return in Procgen. Median and interquartile range over three seeds. The vertical dotted lines separate the pretraining phase from the training phase.

0.79 against 0.62 on DoorKey-16x16 at the phase boundary and 0.76 against 0.72 on KeyCorridor-S6R3 (Figure 4, bottom row), four out of five steps on DoorKey-16x16 reaching a state not yet seen in that episode. *AEGIS* therefore does not explore less; it explores with less redundancy and reaches the terminating state sooner, since a step revisiting an already-seen observation earns nothing (Prop. 1). r^{local} supplies this within-episode signal, while r^{global} supplies direction, steering the agent toward transitions the dynamics model has not yet fitted rather than toward unvisited states as such. Neither alone is sufficient: a policy repeating one trajectory is already optimal for r^{local} (Cor. 1), and r^{global} places no constraint on within-episode diversity (Cor. 2), leaving the agent free to fixate on whichever region currently has the highest prediction error. Only the product enforces both (Theorem 2).

Experiments in Procgen. We evaluate the generalizability of *AEGIS* using three tasks (ninja, climber, jumper) from the Procgen benchmark on hard mode (Cobbe et al., 2020). Unlike Minigrid, Procgen regenerates the level layout at every episode, so the two benchmarks stress different aspects of exploration: Minigrid requires discovering a fixed but long sequence of dependent actions, whereas Procgen presents a fresh configuration each time. For these tasks, we employ a CNN architecture similar to IMPALA for the agent's policy and value functions (Espeholt et al., 2018) and a wider version of Mnih et al. (2015) for the environment's dynamics models. Results are reported in Figure 5.

Two observations follow. First, *AEGIS* matches or exceeds the strongest baseline in every task and budget allocation, so the method transfers from grid worlds to high-dimensional pixel inputs without any change to its hyperparameters. Second, the gap between the two strongest methods and the remaining five widens with the pretraining budget: under NPT all methods lie within a narrow band, whereas under HPT *AEGIS* and DEIR reach the highest return achieved by any method on climber while PPO, ICM, RND, NGU and NovelD remain well below it. The margin between *AEGIS* and DEIR, however, is smaller here than on Minigrid. We attribute this to the source of the exploration difficulty rather than to the input modality: when each episode presents a new layout, procedural variation already exposes the agent to diverse states, and directed exploration adds less on top of it. The advantage of *AEGIS* is largest where the challenge is structural, that is, where the agent must discover a particular ordering of actions in a layout that mostly persists across episodes. We regard this as a scoping statement rather than a limitation of the pixel setting: Procgen shows that *AEGIS* scales, and Minigrid shows where its mechanism pays off most.

5 Ablation Studies

We demonstrate the necessity of *AEGIS*' components through a series of ablation studies on DoorKey-16x16. We initially used DoorKey-8x8, the smaller variant, for its shorter training time, but found that it separates only configurations that fail outright and not those that merely converge more slowly: every variant that learns at all converges within a narrow band there. The component ablations are therefore reported on DoorKey-16x16, which was never used for tuning, and ablations on DoorKey-8x8 are retained in App. B for the coefficient sweeps.

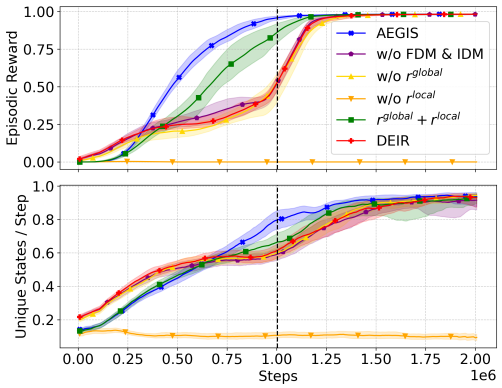


Figure 6: Ablation on r^{int} components.

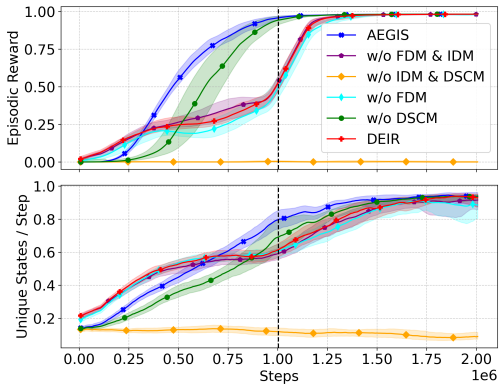


Figure 7: Ablation on $\mathcal{L}^{LMDP}$ components.

Reward components. We test each reward component in isolation, an alternative interaction formula (additive instead of multiplicative), and the loss-side variant that removes the forward and inverse models, with DEIR as a reference (Figure 6). Removing r^{local} is catastrophic: the agent never completes a single episode in any of the ten seeds, and its per-step novelty stays flat for the entire run, the behavior Cor. 2 describes: an agent free to fixate on whichever transitions currently carry high prediction error without any pressure toward within-episode diversity. Removing r^{global} instead returns performance to the DEIR baseline, as does removing $\mathcal{L}^{FDM}$ and $\mathcal{L}^{IDM}$ while keeping both rewards: neither component recovers the full method on its own, the correspondence Remark 8 describes. The additive variant outperforms every single-component ablation, confirming both rewards contribute, but remains clearly behind the multiplicative form throughout pretraining: Proposition 1 explains the gap, since under multiplication a revisited step contributes nothing, whereas under addition it still collects r_t^{global}, so the cycling policy of Cor. 2 is no longer excluded.

Representation components. Figure 7 tests the loss components. Retaining $\mathcal{L}^{FDM}$ alone leaves the objective with its only term that a collapsed encoder can minimize, and the result is the predicted failure: zero return and flat per-step novelty for the whole run, matching Lemma 1. Removing $\mathcal{L}^{FDM}$ while keeping the other two leaves a non-degenerate but geometrically unconstrained latent space, on which the distances entering r^{local} carry no guarantee, and performance falls to the DEIR baseline, which is what Lemma 2 anticipates for an objective built only from metric-blind terms. Removing $\mathcal{L}^{DSCM}$ is the one variant that is not uniformly worse: it lags substantially through early pretraining and then converges marginally after the full objective. We report this rather than tune around it; the discriminator accelerates early exploration at a small cost near convergence. Further weight combinations are reported on DoorKey-8x8 in App. B without significant differences, which is expected, since the weights do not enter the comparison in Theorem 1.

Optimization objective and noisy observations. We also compare the approximate bilevel Stackelberg formulation against min-max (zero-sum) alternatives, and test robustness to the noisy-TV problem by adding Gaussian noise to observations (App. B, Figures 8, 9). All three zero-sum variants fail to learn, with per-step novelty *declining* rather than failing to rise, confirming that the general-sum formulation is not a stylistic preference but the reason the representation does not collapse (Lemma 1); App. B.5 confirms this at the representation level. *AEGIS* is unaffected by noisy observations, since Proposition 1 drives r^{local} to zero once no new observations are found; NovelD's novelty difference over an RND error has no such episodic factor, so noise sustains it indefinitely.

6 Related Work

RFPT methods learn behavior before any task reward is available, falling broadly into: i) skill discovery which rewards behavioral distinguishability rather than reach and so explores locally (as in VALOR (Achiam et al., 2018), DADS (Sharma et al., 2019), DIAYN (Eysenbach et al., 2018) and LSD (Park et al., 2022)), and ii) coverage maximization which addresses reach directly but relies on visitation estimates that are expensive to maintain in high-dimensional observation spaces (as in CBB (Bellemare et al., 2016), MaxEnt (Hazan et al., 2019), SMM (Lee et al., 2019), APT (Liu & Abbeel, 2021), Proto-RL (Yarats et al., 2021) and RE3 (Seo et al., 2021)). Neither family treats representation as driven by the exploration it uses, which is the interdependence *AEGIS* targets.

Exploration-focused methods designed for hard exploration problems navigate expansive state spaces and produce diverse behavior by adding curiosity-driven intrinsic motivation to the environment's often sparse reward, which makes them well suited to RFPT settings. Examples include ICM (Pathak et al., 2017), RND (Burda et al., 2019), Disagreement (Pathak et al., 2019), QXplore (Simmons-Edler et al., 2019), RIDE (Raileanu & Rocktäschel, 2020), Plan2Explore (Sekar et al., 2020), NGU (Badia et al., 2020), NovelD (Zhang et al., 2021) and DEIR (Wan et al., 2023). Three of these are close enough to *AEGIS* to warrant a precise comparison. NGU also multiplies an episodic term by a lifelong one, but computes the episodic term by k-nearest-neighbor lookup in a stored buffer, whereas *AEGIS* computes both from the same learned latent space with no episodic memory at inference. NovelD also uses a novelty *difference*, but over an RND error, which constrains no geometry; *AEGIS* takes it over a latent forward model shaped by an inverse model and a discriminator. DEIR supplies the closest local signal, a discriminative episodic reward, and is our strongest baseline throughout. The distinction that matters (Section A) is two-sided: a discriminative objective is metric-blind (Lemma 2), so it cannot constrain the Euclidean distances standing in for state distance in the denominator of Eq. (7), while a forward model supplies exactly that constraint but allows a collapsed solution (Lemma 1). Adding the forward and inverse models to a discriminative objective is still not sufficient: an episodic reward gives the policy no reason to seek the states on which the model is weak, the coupling Remark 8 identifies. Section 5 reports both halves of that decomposition.

Adversarial RL incorporates game-theoretic logic into RL training in various ways, including perturbing the policy through an adversary for robustness (Fickinger et al., 2021; Oikarinen et al., 2021) and using GANs for online imitation learning (L'Erario et al., 2024). These formulations are typically zero-sum. *AEGIS* departs from that convention deliberately towards a general-sum leader–follower pair, and Lemma 1 shows why. To our knowledge, *AEGIS* is the first method whose components are justified by which degenerate solution each excludes, not by ablations alone.

7 Limitations & Future work

Scope of the theory. Our results characterize which degenerate solutions each component excludes; they are not performance guarantees. We do not claim a bound on the steps needed to cover the state space, and Theorem 1 rules out a degenerate latent geometry without establishing that the geometry approximates reachability (Remark 2). The results also concern the objectives rather than the optimizer: excluding a degenerate global minimizer does not preclude gradient descent lingering near it, and the reward-side results are stated at fixed model parameters. App. A assumes deterministic transitions and emissions, holding in the grid worlds we analyze but not in Procgen; the arguments extend to the stochastic case by replacing trajectory returns with expectations, though we do not carry this out. Assumption 1 requires a representable non-collapsed encoder with zero forward-prediction error: App. A.3 reports the attained values but does not establish representability in general.

Empirical scope. The experiments cover three Minigrid and three Procgen tasks with discrete action spaces; extending to continuous control would require a shift to an off-policy algorithm, which we leave untested. The Procgen results use three seeds against ten for Minigrid, so the smaller margins reported there should be read as indicative rather than conclusive. As discussed in Section 4, the advantage of *AEGIS* is largest where the exploration challenge is structural, and narrower where procedural variation already diversifies the observations the agent encounters.

8 Conclusion

In this work, we identify a critical gap in RFPT for RL: the failure to jointly optimize exploration and representation learning, which leads to brittle generalization and poor downstream performance. To address this, we proposed *AEGIS*, an algorithm that improves representations by targeting their weaknesses, both locally via novelty-seeking and globally via adversarial exploration. Rather than justifying this composition empirically, we showed it follows a single structural principle, instantiated at both the representation and reward level: each component optimizes a proxy objective with a degenerate solution of its own, and each pairing excludes exactly the solution its partner tolerates. Empirically, *AEGIS* reaches near-optimal return substantially sooner than the strongest baselines on Minigrid, with shorter and less redundant trajectories, and transfers to pixel-based Procgen tasks without re-tuning. Its ablations follow the ordering the theory predicts. Our empirical results highlight the benefits of co-optimizing exploration and representation learning during pretraining.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms, 2018.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sye57xStvB.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Nicolò Botteghi, Mannes Poel, and Christoph Brune. Unsupervised representation learning in deep reinforcement learning: A review. *CoRR*, abs/2208.14226, 2022. doi: 10.48550/ARXIV.2208.14226. URL https://doi.org/10.48550/arXiv.2208.14226.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 2048–2056. PMLR, 2020. URL http://proceedings.mlr.press/v119/cobbe20a.html.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1665–1674. PMLR, 2019. URL http://proceedings.mlr.press/v97/du19b.html.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1406–1415. PMLR, 2018. URL http://proceedings.mlr.press/v80/espeholt18a.html.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Arnaud Fickinger, Natasha Jaques, Samyak Parajuli, Michael Chang, Nicholas Rhinehart, Glen Berseth, Stuart Russell, and Sergey Levine. Explore and control with adversarial surprise. *CoRR*, abs/2107.07394, 2021. URL https://arxiv.org/abs/2107.07394.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR, 2019. URL http://proceedings.mlr.press/v97/hazan19a.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Inf. Fusion*, 85:1–22, 2022. doi: 10.1016/J.INFFUS.2022.03.003. URL https://doi.org/10.1016/j.inffus.2022.03.003.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric P. Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *CoRR*, abs/1906.05274, 2019. URL http://arxiv.org/abs/1906.05274.

Giuseppe L'Erario, Drew Hanover, Angel Romero, Yunlong Song, Gabriele Nava, Paolo Maria Viceconte, Daniele Pucci, and Davide Scaramuzza. Learning to walk and fly with adversarial motion priors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pp. 10370–10377. IEEE, 2024. doi: 10.1109/IROS58592.2024.10802391. URL https://doi.org/10.1109/IROS58592.2024.10802391.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

N. Nisan, T. Roughgarden, E. Tardos, and V.V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007. ISBN 9780521872829. URL https://books.google.ie/books?id=f4SFngEACAAJ.

Tuomas P. Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 26156–26167, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/dbb422937d7ff56e049d61da730b3e11-Abstract.html.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pp. 5062–5071. PMLR, 2019.

Roberta Raileanu and Tim Rocktäschel. RIDE: rewarding impact-driven exploration for procedurally-generated environments. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkg-TJBFPB.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pp. 8583–8592. PMLR, 2020.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pp. 9443–9454. PMLR, 2021.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Riley Simmons-Edler, Ben Eisner, Eric Mitchell, H. Sebastian Seung, and Daniel D. Lee. Qxplore: Q-learning exploration by maximizing temporal difference error. *CoRR*, abs/1906.08189, 2019. URL http://arxiv.org/abs/1906.08189.

Matthijs TJ Spaan. Partially observable markov decision processes. In *Reinforcement learning: State-of-the-art*, pp. 387–414. Springer, 2012.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9870–9879. PMLR, 2021. URL http://proceedings.mlr.press/v139/stooke21a.html.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shanchuan Wan, Yujin Tang, Yingtao Tian, and Tomoyuki Kaneko. DEIR: efficient and robust exploration through discriminative-model-based episodic intrinsic rewards. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 4289–4298. ijcai.org, 2023. doi: 10.24963/IJCAI.2023/477. URL https://doi.org/10.24963/ijcai.2023/477.

Zhihui Xie, Zichuan Lin, Junyou Li, Shuai Li, and Deheng Ye. Pretraining in deep reinforcement learning: A survey. *arXiv preprint arXiv:2211.03959*, 2022.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pp. 11920–11931. PMLR, 2021.

Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E. Gonzalez, and Yuandong Tian. Noveld: A simple yet effective exploration criterion. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 25217–25230, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d428d070622e0f4363fceae11f4a3576-Abstract.html.

AI USE STATEMENT

In this work, we used generative AI tools to [*provide critical ingredients for proving mathematical claims, formulate mathematical claims, assist in the writing of proofs*].

We have not used generative AI tools to [*implement methods, assist with translation, interpret results, design or provide feedback on research methodology or experiments*],

and [*Generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis*] are not applicable to this work.

Additionally, we used generative AI tools to [*create or edit software code, identify relevant literature, edit a research paper to improve readability*].

We have reviewed all AI-assisted work. The code created was limited to plotting and figure formatting, not the method or experiments themselves. Text edits to improve readability were always verified by the authors. In our mathematical claims, AI assistance involved standardizing notations to ensure consistency, providing critical scrutiny, and text edits to improve readability and understanding.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies exploration and representation learning in simulated grid-world and video-game benchmarks. It involves no human subjects, no personal data, and no newly released datasets. We do not foresee direct societal harms arising from the method itself, though improvements to sample-efficient reinforcement learning carry the general dual-use considerations of the field.

REPRODUCIBILITY STATEMENT

The algorithmic description of *AEGIS* is given in Section 3 and Algorithm 1. App. A states all theoretical results, with assumptions made explicit and proofs complete. Environment descriptions, network architectures, computing environment, and the full hyperparameter tables for *AEGIS* and every baseline are in App. C. All reported curves use ten seeds on Minigrid and three on Procgen, with medians and interquartile ranges as described in Section 4. We will make the source code available on GitHub.

A THEORETICAL JUSTIFICATION OF THE COMPONENTS

Both the representation model and the intrinsic reward of *AEGIS* are built from several components. A natural objection is that such compositions are assembled empirically rather than derived. In this section we show that the composition follows a single structural principle: *Each individual component optimizes a proxy objective that has a degenerate solution when used in isolation, and each pairing eliminates exactly the degenerate solution that the other component tolerates.*

We instantiate this principle twice: once at the representation level (Section A.1) and once at the reward level (Section A.2). We characterize which failure modes each component provably excludes, and each result corresponds to an ablation already reported in Section 5.

Notation. We keep the notation of Sections 2 and 3: f_θ is the LMDP encoder and we write $f_\theta = (f_\theta^{obs}, f_\theta^{traj})$ for its CNN and RNN components, with outputs $z^{obs} \in \mathcal{Z}^{obs}$ and $z^{traj} \in \mathcal{Z}^{traj} = \mathbb{R}^d$. The heads are the forward model h_δ, the inverse model q_ζ and the discriminator g_ω, all three operating on z^{traj}, with losses $\mathcal{L}^{FDM}$, $\mathcal{L}^{IDM}$, $\mathcal{L}^{DSCM}$ as in Eqs. (2)–(4). We write θ for the LMDP parameters collectively. Three further conditions are used below: i) $|\mathcal{A}| \geq 2$ under a non-deterministic behavior policy, so that the conditional action entropy $\mathbb{H}(a|t) > 0$; ii) the actor conditions on z_t^{traj}, so the policy class is history-dependent rather than stationary, and iii) f_θ is injective on the observations visited within a single episode, which the block property and point-mass emissions make attainable.

A.1 THE REPRESENTATION MODEL

Definition 1 (Collapsed encoder)**.** The encoder is *observation-collapsed* if there is a $c \in \mathcal{Z}^{obs}$ with $f_\theta^{obs}(o) = c$ for every observation in the support of the data distribution, and *trajectory-collapsed* if there is a $c \in \mathcal{Z}^{traj}$ with $f_\theta^{traj}(h) = c$ for every history in that support. It is *collapsed* if it is collapsed in either sense. An objective $\mathcal{L}$ *allows collapse* if some collapsed f_θ lies in $\arg\min_\theta \mathcal{L}(\theta)$, and *excludes collapse* if every collapsed f_θ satisfies $\mathcal{L}(\theta) > \min_\theta \mathcal{L}(\theta)$.

Definition 2 (Affine reparametrization and metric blindness)**.** Let $\varphi(z) = Az + b$ with $A \in \mathbb{R}^{d\times d}$ invertible, acting on $\mathcal{Z}^{traj}$, the space the LMDP heads consume. An objective $\mathcal{L}$ is *metric-blind* if for every encoder f_θ and every such φ there exist head parameters attaining $\mathcal{L}(\varphi \circ f_\theta) = \mathcal{L}(f_\theta)$; otherwise it is *metric-sensitive*.

Metric blindness matters because the denominator of r^{local} (Eq. (7)) is a Euclidean distance in $\mathcal{Z}^{traj}$, standing in for the distance between the underlying states (Section 3). If the representation objective cannot distinguish a latent geometry from an arbitrarily distorted copy of it, then it places no constraint on that term. The numerator is a distance in $\mathcal{Z}^{obs}$, and what it requires is not a constrained metric but that distinct observations receive distinct embeddings, which is the separation property of Lemma 1. The two halves of Eq. (7) are therefore constrained by different components.

Lemma 1 compares the value a collapsed encoder attains against the best value attainable in the model class, so it presupposes that a non-degenerate encoder is representable.

Assumption 1. There exists a non-collapsed θ^* with $\mathcal{L}^{FDM}(\theta^*) = 0$, $\mathcal{L}^{IDM}(\theta^*) = \eta_I < \mathbb{H}(a \mid t)$ and $\mathcal{L}^{DSCM}(\theta^*) = \eta_D < \ln 2$, where η_I and η_D denote the attained values.

Subsection A.3 reports the attained values of $\mathcal{L}^{FDM}$, η_I and η_D against their bounds for *AEGIS* on DoorKey-16x16.

Assumption 1 holds whenever transitions are deterministic, actions are identifiable from consecutive latent states, and the encoder class is expressive enough to represent the block map $o \mapsto s$: the FDM then predicts the next latent exactly, while the IDM and DSCM operate strictly better than a timestep-conditioned prior and than chance, respectively.

Lemma 1 (Representation Collapse)**.** *$\mathcal{L}^{FDM}$ allows collapse. $\mathcal{L}^{IDM}$ and $\mathcal{L}^{DSCM}$ exclude it.*

Proof of Lemma 1.
(i) FDM. $\mathcal{L}^{FDM} = \mathcal{J}^{MSE}\Big(h_\delta(z_t^{traj}, a_t),\, z_{t+1}^{traj}\Big) \geq 0$, so $\min_\theta \mathcal{L}^{FDM} \geq 0$. Take $f_\theta^{traj} \equiv c$ and any h_δ with $h_\delta(c, a) = c$ for all $a \in \mathcal{A}$ (realizable, e.g. by zeroing the output layer weights and

setting its bias to c). Then every prediction equals its target and $\mathcal{L}^{FDM}(\theta) = 0$, hence a collapsed encoder lies in $\arg\min_\theta \mathcal{L}^{FDM}(\theta)$.

(ii) IDM. Under trajectory collapse the input (c, c) to q_ζ is constant, so q_ζ induces a single distribution over $\mathcal{A}$ independent of the transition. The categorical cross-entropy is minimized over such constant predictors by the marginal action distribution, giving $\mathcal{L}^{IDM}(\theta) = \mathbb{H}(a)$. Under observation collapse, $f_\theta^{obs} \equiv c$, so the RNN receives a constant input stream and $z_t^{traj} = g(t)$ for a deterministic g of the step index alone. The input $(g(t), g(t+1))$ then carries no information about a_t beyond t itself, and the best predictor is $p(a \mid t)$, giving $\mathcal{L}^{IDM}(\theta) \geq \mathbb{H}(a \mid t)$. Since $\mathbb{H}(a \mid t) \leq \mathbb{H}(a)$, every collapsed encoder attains at least $\mathbb{H}(a \mid t)$. By Assumption 1 there is a non-collapsed θ^* with $\mathcal{L}^{IDM}(\theta^*) = \eta_I < \mathbb{H}(a \mid t)$, so $\min_\theta \mathcal{L}^{IDM}(\theta) < \mathbb{H}(a \mid t)$ and no collapsed encoder lies in $\arg\min_\theta \mathcal{L}^{IDM}$.

(iii) DSCM. Under trajectory collapse the discriminator input is (c, c, a_t) for both the positive o_{t+1}^+ and the negative o^-. Under observation collapse, the positive and the negative observation map to the same $z^{obs} = c$, so the RNN returns the same z_{t+1}^{traj} in either case and the input is again identical. In both cases, positives and negatives are indistinguishable, and the optimal constant predictor outputs the class prior. The objective weights the positive and the negative term equally (Section 3), so that the prior is $1/2$ and the attained binary cross-entropy is $\ln 2 > 0$. By Assumption 1, $\min_\theta \mathcal{L}^{DSCM}(\theta) \leq \eta_D < \ln 2$, so no collapsed encoder lies in $\arg\min_\theta \mathcal{L}^{DSCM}$. □

Lemma 2 (Metric blindness). *$\mathcal{L}^{IDM}$ and $\mathcal{L}^{DSCM}$ are metric-blind. $\mathcal{L}^{FDM}$ is metric-sensitive.*

Proof of Lemma 2.
(i) IDM, DSCM. Both heads consume latents only as inputs to an MLP whose first layer is affine: q_ζ takes $(z_t^{traj}, z_{t+1}^{traj})$ and g_ω takes $(z_t^{traj}, z_{t+1}^{traj}, a_t)$. Replacing f_θ by $\varphi \circ f_\theta$ and composing the first layer of the head with φ^{-1} on the latent coordinates leaves every head *output* unchanged; since the targets (a_t and the binary label) live outside $\mathcal{Z}^{traj}$ and are unaffected by φ, both losses are unchanged. The modified head lies in the same parametric family, as composing an affine first layer with an invertible affine map yields an affine map. Thus, IDM and DSCM are metric-blind.

(ii) FDM. Here the target is itself a latent vector. Take $\varphi(z) = \lambda z + b$, $\lambda \notin \{0, 1, -1\}$: the first exclusion makes φ invertible and hence a member of the family in Definition 2, while the other two exclude the isometries, under which $\mathcal{L}^{FDM}$ is invariant (Remark 1). So $\varphi^{-1}(z) = (z - b)/\lambda$, and we write the per-transition error of (f_θ, h_δ) as $e = h_\delta(z_t^{traj}, a_t) - z_{t+1}^{traj}$, so that $\mathcal{L}^{FDM}(f_\theta) = \mathbb{E}\big[\|e\|^2\big]$.

Since metric-blindness asserts invariance under every such map, exhibiting a single φ that changes the loss establishes metric-sensitivity. Under the warped encoder $\varphi \circ f_\theta$ the head receives $\lambda z_t^{traj} + b$ and is compared against the warped target $\lambda z_{t+1}^{traj} + b$, so the transported head $h'_\delta = \varphi \circ h_\delta \circ (\varphi^{-1} \times \mathrm{id})$ must undo the warp on its input and reapply it on its output; unlike the IDM and DSCM, whose targets lie outside $\mathcal{Z}^{traj}$, this second map cannot be absorbed. Writing this out:

$$
\begin{aligned}
h'_\delta\big(\lambda z_t^{traj} + b,\, a_t\big) &= \varphi\Big(h_\delta\big(\varphi^{-1}(\lambda z_t^{traj} + b),\, a_t\big)\Big) = \lambda\, h_\delta\big(z_t^{traj}, a_t\big) + b, \\
e' &= \big(\lambda\, h_\delta(z_t^{traj}, a_t) + b\big) - \big(\lambda\, z_{t+1}^{traj} + b\big) = \lambda e, \\
\mathcal{L}^{FDM}(\varphi \circ f_\theta) &= \mathbb{E}\big[\|\lambda e\|^2\big] = \lambda^2\, \mathcal{L}^{FDM}(f_\theta),
\end{aligned}
$$

using $\varphi^{-1} \circ \varphi = \mathrm{id}$ in the first line, the scaled target $\varphi(z_{t+1}^{traj}) = \lambda z_{t+1}^{traj} + b$ in the second, and $\|\lambda v\| = |\lambda|\, \|v\|$ in the third. The map $h_\delta \mapsto h'_\delta$ is a bijection on the head family, so taking minima over heads gives $\min_\delta \mathcal{L}^{FDM}(\varphi \circ f_\theta) = \lambda^2 \min_\delta \mathcal{L}^{FDM}(f_\theta)$: the ratio holds for the best head, not merely for one particular choice. Whenever $\min_\delta \mathcal{L}^{FDM}(f_\theta) > 0$ the two values differ, so no head parameters preserve the loss and the objective is metric-sensitive. □

Remark 1 (What $\mathcal{L}^{FDM}$ pins down). The existential form of Definition 2 is necessary rather than merely convenient. Writing e for the per-transition error as above, a general $\varphi(z) = Az + b$ gives $e' = Ae$, since the translation shifts prediction and target alike, and hence $\mathcal{L}^{FDM}(\varphi \circ f_\theta) = \mathbb{E}\big[\|Ae\|^2\big]$. This equals $\mathbb{E}\big[\|e\|^2\big]$ for every encoder exactly when $A^\top A = I$; for a particular error distribution a non-orthogonal A may happen to preserve the value, but only the orthogonal maps do so universally. The loss is invariant under the isometries of $\mathcal{Z}^{traj}$, the rotations, reflections and translations that preserve distances, and changes under every map that distorts them. No single φ

could therefore establish invariance-breaking for all of them, which is why metric-sensitivity is stated existentially; the scaling $\varphi(z) = \lambda z + b$ used above is simply the most convenient member of the distance-distorting class. One caveat is worth recording. At an encoder attaining $\mathcal{L}^{FDM} = 0$, such as the θ^* of Assumption 1, the scaling argument gives $\lambda^2 \cdot 0 = 0$, so every affine copy attains the same value and the objective does not distinguish them. The metric constraint is therefore informative in the regime of imperfect prediction, which is the realistic one and the one in which r^{local} operates, and this is a further reason the positive alignment result is deferred.

Consequently, $\mathcal{L}^{FDM}$ determines the latent geometry *up to isometry*: it leaves undetermined only the orientation and origin of $\mathcal{Z}^{traj}$, and constrains everything about the distances between embeddings. This is the appropriate level of determination for our purposes, since r^{local} (Eq. 7) is a function of Euclidean distances and is itself isometry-invariant, so the residual freedom does not affect the reward. By contrast, $\mathcal{L}^{IDM}$ and $\mathcal{L}^{DSCM}$ are invariant under the full affine group, including maps that distort distances arbitrarily, and so leave the quantity r^{local} consumes entirely unconstrained. We stress that this concerns sensitivity of the objective, not the quality of the solution: that minimizing $\mathcal{L}^{FDM}$ yields distances approximating reachability is the separate alignment question deferred in Remark 2.

Lemmas 1 and 2 already show that the two properties are disjoint across components: $\mathcal{L}^{FDM}$ is the only metric-sensitive term and the only one that collapse can minimize, so no single component provides both. What does not follow immediately is that either property survives the weighted sum of Eq. (1), since a sum of objectives need not inherit the minimizer of its terms. Theorem 1 establishes this.

Theorem 1. *Let $\mathcal{L}^{LMDP} = w_F\mathcal{L}^{FDM} + w_I\mathcal{L}^{IDM} + w_D\mathcal{L}^{DSCM}$ with $w_F, w_I, w_D > 0$ as in Eq. (1), and let Assumption 1 hold. Then $\mathcal{L}^{LMDP}$ excludes collapse and is metric-sensitive, although by Lemmas 1–2 no single component has both properties.*

Proof of Theorem 1. Collapse. Let θ_c be any collapsed encoder. By Lemma 1, $\mathcal{L}^{FDM}(\theta_c) \geq 0$, $\mathcal{L}^{IDM}(\theta_c) \geq \mathbb{H}(a \mid t)$ and $\mathcal{L}^{DSCM}(\theta_c) \geq \ln 2$, so

$$\mathcal{L}^{LMDP}(\theta_c) \;\geq\; w_I\,\mathbb{H}(a \mid t) + w_D \ln 2.$$

For θ^* from Assumption 1: $\mathcal{L}^{LMDP}(\theta^*) = w_I\,\eta_I + w_D\,\eta_D$. Since $w_I, w_D > 0$, $\eta_I < \mathbb{H}(a \mid t)$ and $\eta_D < \ln 2$, we obtain $\mathcal{L}^{LMDP}(\theta^*) < \mathcal{L}^{LMDP}(\theta_c)$, so $\theta_c \notin \arg\min_\theta \mathcal{L}^{LMDP}(\theta)$, which means no collapsed encoder is a global minimizer of $\mathcal{L}^{LMDP}$. Note that the $\mathcal{L}^{FDM}$ term cannot compensate: it is non-negative, so dropping it from the lower bound is valid and its weight w_F does not enter the comparison, making the conclusion independent of the choice of weights.

Metric sensitivity. Take $\varphi(z) = \lambda z + b$ with $\lambda \notin \{0, 1, -1\}$, the map used in Lemma 2. By that lemma the IDM and DSCM terms are unchanged under φ, while the FDM term scales by λ^2, so

$$\mathcal{L}^{LMDP}(\varphi \circ f_\theta) - \mathcal{L}^{LMDP}(f_\theta) = w_F\,(\lambda^2 - 1)\,\mathcal{L}^{FDM}(f_\theta),$$

which is non-zero whenever $\mathcal{L}^{FDM}(f_\theta) > 0$. Hence the composite objective is not invariant under affine reparameterization of $\mathcal{Z}^{traj}$. □

Remark 2. Theorem 1 states that the composite objective *rules out* a degenerate latent geometry on $\mathcal{Z}^{traj}$; we recognize the limitation that this does not establish that the resulting Euclidean distances approximate reachability. A positive alignment result of that kind is left to future work.

A.2 The Intrinsic Reward

Recall from Section 2 that transitions and emissions are deterministic in the environments we evaluate; the arguments extend to the stochastic case by replacing trajectory returns with their expectations. For an intrinsic reward r we write $R_r(\tau) = \sum_{t=0}^{H-1} r_t$ for the return of a single episode following τ, $J_r(\pi) = \mathbb{E}_\pi[R_r(\tau)]$ for the expected per-episode return, and α for the intrinsic reward coefficient of Eq. (5).

This subsection proceeds in two steps. We first show that each reward component, taken alone and at fixed model parameters, has a globally optimal policy that fails to explore, and that the two failures are of different kinds. We then show that the product excludes both, and that what it rewards instead is *non-redundant* behavior: a step that returns to an observation already seen in the episode earns nothing.

To state the two failures, we use two counting quantities. The *per-episode* coverage is $\mathcal{C}(\pi) = \mathbb{E}_\pi|\mathcal{S}(\tau)|$ and the *cumulative* coverage over episodes 1 through E is $\mathcal{C}_{1:E}(\pi) = \mathbb{E}_\pi\big|\bigcup_{e=1}^{E}\mathcal{S}(\tau_e)\big|$, the number of distinct states visited at least once within an episode and within the first E episodes respectively. For the positive result, we use the *per-episode redundancy*:

$$\rho(\pi) = \mathbb{E}_\pi\left[\frac{|\mathcal{S}(\tau)| - 1}{|\tau|}\right] \in [0, 1], \tag{11}$$

the fraction of steps that reach a state not yet visited in the episode; the initial state is excluded from the count since no step reaches it. We report this quantity as distinct states per step in Section 4.

Definition 3 (Coverage blindness)**.** Let h_t^{ep} denote the within-episode history up to step t, and θ the model parameters, fixed while the policy is optimized within an update cycle. A reward is *inter-episode blind* if it carries no information about other episodes: $r_t = r(h_t^{ep}; \theta)$, and *visitation-blind with window* k if it carries no information about states visited earlier in the episode: $r_t = r\big(\{(o_j, a_j, o_{j+1})\}_{j=t-k+1}^{t}; \theta\big)$. In both cases, θ may encode accumulated training knowledge; what is excluded is dependence on earlier episodes and the within-episode visitation set, respectively.

The two notions are not exclusive. r^{global} depends only on the transitions at $t-1$ and t and on θ, so it is visitation-blind with $k = 2$ and inter-episode blind, and Lemma 3 applies to it as well as to r^{local}. What separates the two components is the converse: r^{local} is a minimum over the entire within-episode history, so it is not visitation-blind for any finite k.

Lemma 3 (Inter-episode blind rewards have low-coverage optima)**.** *If r is inter-episode blind, there exists a globally optimal policy $\pi^\star$ for J_r that executes a single fixed trajectory $\tau^\star$ in every episode, so $\mathcal{C}_{1:E}(\pi^\star) = |\mathcal{S}(\tau^\star)| \le H + 1$ for every E, irrespective of $|\mathcal{S}|$.*

Proof of Lemma 3. Inter-episode blindness makes the episodic return a function of the trajectory alone, $R_r(\tau) = \sum_{t=0}^{H-1} r_t(h_t^{ep})$. The set of feasible trajectories is finite ($|\mathcal{S}|, |\mathcal{A}|, H < \infty$), so $\tau^\star \in \arg\max_\tau R_r(\tau)$ exists. Let $\pi^\star$ be the policy realizing $\tau^\star$; such a policy exists in the history-dependent class used by AEGIS even when $\tau^\star$ visits a state twice under different actions, which a stationary policy could not reproduce. For any policy π, every episode it generates has return at most $R_r(\tau^\star)$, hence $J_r(\pi) \le R_r(\tau^\star) = J_r(\pi^\star)$ and $\pi^\star$ is globally optimal. Since $\pi^\star$ realizes the same $\tau^\star$ in every episode, the union in $\mathcal{C}_{1:E}$ is $\mathcal{S}(\tau^\star)$ for every E, and since R_r does not depend on the episode index, no repetition of $\tau^\star$ reduces its return. □

Corollary 1 (r^{local} alone does not expand coverage across episodes)**.** *At any fixed encoder parameter θ, r^{local} of Eq. (7) is inter-episode blind, so by Lemma 3 a policy repeating one fixed trajectory $\tau^\star$ is globally optimal for $J_{r^{local}}$ and satisfies $\mathcal{C}_{1:E}(\pi^\star) = |\mathcal{S}(\tau^\star)|$ for every E: no state outside $\mathcal{S}(\tau^\star)$ is visited, however many episodes it runs.*

Proof of Corollary 1. r_t^{local} is a minimum over the embeddings z_i^{obs}, z_i^{traj} with $i \in [0, t)$ accumulated in the current episode. These histories are initialized at episode start and cleared at episode end, and no other term of Eq. (7) depends on data outside the episode. Hence r_t^{local} is a function of h_t^{ep} and of θ, which is Definition 3. Apply Lemma 3. □

Corollary 2 (r^{global} alone does not enforce within-episode diversity)**.** *At any fixed LMDP parameter θ, r^{global} of Eq. (9) is visitation-blind with window $k = 2$. Moreover, in any environment containing a cycle $C \ni s_0$ of length $m \le H$ there exist parameter settings at which the policy traversing C is globally optimal for $J_{r^{global}}$, giving $\mathcal{C}(\pi) = m$ independently of $|\mathcal{S}|$. The second part is an existence statement: it exhibits parameter settings at which visitation blindness makes such a policy optimal. We do not show that training reaches these settings, though Remark 3 describes why configurations of this kind arise transiently, and Section 5 reports the corresponding behavior.*

Proof of Corollary 2. $r_t^{global} = \max\big(v_t - 0.5\, v_{t-1}, 0\big) + 1$ with v_t the per-transition FDM error of Eq. (8) depends only on the transitions at t and $t-1$ and on θ; it contains no term depending on the set of states already visited in the episode, which is Definition 3 with $k = 2$. For the construction, let θ be such that $v = V > 0$ on the transitions of C and $v = 0$ elsewhere, with the convention $v_{-1} = 0$. The per-step reward has four cases (Table 2). Cases 1 and 2 both yield reward 1, since the clipping removes the negative difference, so the per-step reward takes three distinct values. Every step of case 3 other than $t = 0$ is preceded by a step with $v_{t-1} = 0$, that is a step of case 1 or 2, and distinct

Table 2: Global reward calculations based on transition cases.

Case	v_t	v_{t-1}	$v_t - 0.5v_{t-1}$	reward
1	0	0	0	1
2	0	V	$-0.5V < 0 \rightarrow$ clipped to 0	1
3	V	0	V	$V+1$
4	V	V	$0.5V$	$0.5V+1$

case-3 steps have distinct predecessors. Writing n_1, n_2, n_3 and n_4 for the number of steps in each case, we have $n_1 + n_2 + n_3 + n_4 = H$, and the pairing gives $n_3 \leq n_1 + n_2 + 1$, the $+1$ coming from $t = 0$ via the convention $v_{-1} = 0$. The total return is:

$$\begin{aligned} R_{r^{global}}(\tau) &= n_1 + n_2 + n_3(V+1) + n_4(0.5V+1) = \\ &= H + Vn_3 + 0.5Vn_4 = H + 0.5Vn_3 + 0.5Vn_3 + 0.5Vn_4 = \\ &= H + 0.5V(n_3 + n_4) + 0.5Vn_3 = H + 0.5V(H - n_2 - n_1) + 0.5Vn_3 = \\ &= H + 0.5VH + 0.5Vn_3 - 0.5Vn_2 - 0.5Vn_1 = \\ &= H(0.5V+1) + 0.5V(n_3 - n_2 - n_1) \end{aligned}$$

which is maximized exactly when $n_3 - n_2 - n_1$ is largest, that is when $n_3 = n_2 + n_1 + 1$. The policy traversing C attains this with $n_1 = 0$, $n_2 = 0$, $n_3 = 1$ and $n_4 = H - 1$, so it is globally optimal, with $\mathcal{C}(\pi) = m$. Other policies also attain $n_3 - n_2 - n_1 = 1$; the claim is that a low-coverage optimum exists, not that it is the only optimum. □

Remark 3 (Scope of Corollary 2). Corollary 2 is stated at fixed θ. It does not claim that the cycling persists indefinitely: revisiting a transition reduces its FDM error, so r^{global} decays where the agent has already been, and the cycle is eventually abandoned. The result is exploration by successive local fixations rather than broad within-episode coverage. No comparable mechanism erodes r^{local}, whose drift with the encoder geometry is uncoupled from how often $\tau^\star$ has been repeated.

Proposition 1 (Revisits earn nothing)**.** *Let $r^{int} = \alpha\, r^{local} r^{global}$ as in Eq. (5), and let π be any policy. If at step t the observation o_{t+1} equals some o_i with $i \in [0, t)$ in the same episode, then $r_t^{local} = 0$ and hence $r_t^{int} = 0$, whatever value r_t^{global} takes.*

Proof of Proposition 1. Under deterministic emissions, two visits to the same latent state produce the same observation and hence the same embedding z^{obs}. If $o_{t+1} = o_i$ for some $i \in [0, t)$, the episodic history contains an entry identical to the current one, so the corresponding term $d^2(z_i^{obs}, z_{t+1}^{obs})/(d(z_i^{traj}, z_t^{traj}) + \epsilon)$ has numerator 0 and finite positive denominator, giving value 0. Since r^{local} is the minimum over the history, $r_t^{local} = 0$ and therefore $r_t^{int} = \alpha\, r_t^{local} r_t^{global} = 0$ regardless of r_t^{global}. □

Remark 4 (Reward component multiplication justification). The proof of Proposition 1 also shows why the two components are multiplied rather than added. Under addition $r_t^{int} = \alpha(r_t^{local} + r_t^{global}) \geq \alpha$ even when $r_t^{local} = 0$, because $r_t^{global} \geq 1$ by construction: a revisited state still earns. Proposition 1 would then fail and with it the exclusion of the cycling policy in Theorem 2. The additive variant is tested in Section 5 and performs accordingly.

Proposition 1 needs only deterministic emissions. It requires no constructed environment, no choice of θ, and no assumption on the policy, which distinguishes it from Corollaries 1 and 2. Two consequences follow.

Corollary 3 (The cycling policy earns nothing)**.** *The policy traversing C in Corollary 2 attains $r^{int} = 0$ at every step after its first revisit.*

Proof of Corollary 3. Since $m \leq H$, the policy completes the cycle at least once within the episode, so from that point on every observation it receives has already appeared in the episodic history. Proposition 1 then gives $r_t^{int} = 0$ at each such step. □

Corollary 4 (The return is supported on non-redundant steps)**.** *For any trajectory τ, $R_{r^{int}}(\tau) = \alpha \sum_{t \in N(\tau)} r_t^{local} r_t^{global}$, where $N(\tau)$ is the set of steps reaching a state not yet visited in the episode, $|N(\tau)| = |\mathcal{S}(\tau)| - 1$. A policy therefore collects intrinsic reward on a $\rho(\pi)$ fraction of its steps and nothing on the remainder.*

Proof of Corollary 4. By Proposition 1 every step $t \notin N(\tau)$ contributes $r_t^{int} = 0$ to the sum $R_{r^{int}}(\tau) = \sum_t r_t^{int}$, leaving only the steps in $N(\tau)$. Each state of $\mathcal{S}(\tau)$ other than s_0 is reached for the first time at exactly one step, so $|N(\tau)| = |\mathcal{S}(\tau)| - 1$ and the rewarded fraction of steps is $|N(\tau)|/|\tau|$ for each trajectory. Taking expectations over π gives $\rho(\pi)$ by Eq. (11). □

Remark 5 (Scope of Corollary 4). Corollary 4 describes the reward, not the optimal policy. A step that returns to a visited state earns nothing even when it is a necessary transit, so the intrinsic reward penalizes unavoidable and avoidable revisits alike. The claim is that redundant steps are unrewarded, not that an optimal policy never revisits.

Proposition 2 (The repeating policy is dominated)**.** *Suppose repetition of $\tau^\star$ has driven the FDM error to zero on every one of its transitions. At the resulting parameters $\tau^\star$ need no longer maximize $R_{r^{local}}$; assume there remains a trajectory $\tilde{\tau}$ with $R_{r^{local}}(\tilde{\tau}) \geq R_{r^{local}}(\tau^\star)$ containing a step t at which $v_t > 0.5 v_{t-1}$ and $r_t^{local} > 0$. Then the repeating policy of Corollary 1 is strictly dominated under $J_{r^{int}}$.*

Proof of Proposition 2. By hypothesis $v_t = 0$ on every transition of $\tau^\star$, so $v_t - 0.5\, v_{t-1} = 0$ and $r_t^{global} = 1$, its floor, at every step of $\tau^\star$. The repeating policy therefore collects only its local reward, $J_{r^{int}}(\pi^\star) = \alpha R_{r^{local}}(\tau^\star)$. For $\tilde{\tau}$, every factor satisfies $r_t^{global} \geq 1$ and $r_t^{local} \geq 0$, so replacing each r_t^{global} by its floor can only decrease the sum:

$$J_{r^{int}}(\tilde{\pi}) = \alpha \sum_t r_t^{local}\, r_t^{global} \overset{(1)}{\geq} \alpha R_{r^{local}}(\tilde{\tau}) \overset{(2)}{\geq} \alpha R_{r^{local}}(\tau^\star) = J_{r^{int}}(\pi^\star),$$

where (1) replaces each r_t^{global} by its floor and (2) is the hypothesis on $\tilde{\tau}$. At the step t of the hypothesis $r_t^{global} > 1$ and $r_t^{local} > 0$ hold together, so that step contributes $r_t^{local} r_t^{global} > r_t^{local}$ and (1) is strict. Hence $J_{r^{int}}(\tilde{\pi}) > J_{r^{int}}(\pi^\star)$. □

Theorem 2. *At fixed θ, neither component alone induces both growth of $\mathcal{C}_{1:E}$ and non-trivial $\mathcal{C}$: by Corollary 1, r^{local} has a globally optimal policy whose $\mathcal{C}_{1:E}$ is constant in E, and by Corollary 2, r^{global} allows an optimal policy with $\mathcal{C}(\pi) = m$ for a cycle length independent of $|\mathcal{S}|$. Their product excludes the cycling policy unconditionally by Corollary 3, and the repeating policy under the hypothesis of Proposition 2. What the product rewards in their place is characterized by Corollary 4.*

Proof of Theorem 2. The first claim is Corollary 1, the second Corollary 2. For the product, Corollary 3 gives $r^{int} = 0$ for the cycling policy, which is therefore not optimal: under a non-collapsed encoder (Theorem 1) and the injectivity condition stated in Section A, any policy reaching a state unvisited in the episode has $d(z_i^{obs}, z_{t+1}^{obs}) > 0$ for every entry i of the episodic history, so $r^{local} > 0$ and hence $r^{int} > 0$; Proposition 2 gives $J_{r^{int}}(\tilde{\pi}) > J_{r^{int}}(\pi^\star)$ for the repeating policy under its hypothesis. □

Remark 6 (Intrinsic reward normalization). r^{int} is rescaled by a running mean and standard deviation over the rollout buffer, reset at every update. This is a positive affine map, so it preserves the ordering of returns of equal-length episodes. The policies of Corollaries 1 and 2 both run the full horizon, so none of the results above is affected. In particular, because the statistics are reset at every model update, normalization introduces no persistent cross-episode signal that could substitute for r^{global}.

Proposition 1 requires one clarification, since it is stated as an exact zero. Since $r^{int} \geq 0$, a revisited step's raw value of 0 is the buffer minimum, so after normalization it scores $-\mu/\sigma$, the smallest value any step in the buffer attains. What the proposition asserts is preserved: a revisit contributes the minimum available at that update, and since the map is affine with positive scale, the ordering underlying Corollaries 3 and 4 is unchanged. Only the additive decomposition of Corollary 4 is stated for the unnormalized reward.

Remark 7 (Correspondence with the ablations). Theorem 1 predicts the $\mathcal{L}^{LMDP}$ ablations: omitting $\mathcal{L}^{IDM}$ and $\mathcal{L}^{DSCM}$ removes the only protection against collapse, while omitting $\mathcal{L}^{FDM}$ retains a non-degenerate but geometrically unconstrained latent space on which r^{local} is computed. Since the weights do not enter the collapse comparison, the theorem also predicts the insensitivity to loss coefficients reported in the weight ablation. Theorem 2 predicts the "w/o r^{local}" and "w/o r^{global}" curves, whose failure modes are the two configurations sketched in Figure 1; and Proposition 1 predicts the robustness of AEGIS to the noisy observations, since a policy that stops discovering new observations drives r^{local}, and hence r^{int}, to zero.

Remark 8 (Coupling of the two levels). Theorem 1 and Theorem 2 are each stated with the other level held fixed: the first characterizes the minimizers of $\mathcal{L}^{LMDP}$ at a given data distribution, the second the optima of $J_{r^{int}}$ at given θ. Neither is vacuous, but each presupposes what the other supplies.

Definition 1 defines collapse on the support of the observation distribution, which is $\mathcal{D}(\pi_\phi)$. Theorem 1 therefore excludes collapse only on the states the policy visits, and the breadth of that support is what the intrinsic reward determines. Conversely, r^{global} is a function of the forward-model error, so if $\mathcal{L}^{FDM}$ is absent from Eq. (1) no term of the objective optimizes h_δ; revisiting a transition then no longer reduces its error, the erosion described in Remark 3 does not occur, and the cycling policy of Corollary 2 becomes a persistent rather than a transient optimum.

The representation-level guarantee is thus relative to a support the reward controls, and the reward's non-degeneracy relies on a model term the objective supplies. Section 5 reports the corresponding ablation: removing either level returns performance to the baseline that has neither.

A.3 Empirical verification of Assumption 1

Assumption 1 asserts the existence of a non-collapsed θ^* attaining $\mathcal{L}^{FDM}(\theta^*) = 0$, $\eta_I < \mathbb{H}(a \mid t)$ and $\eta_D < \ln 2$. We check these against the values attained by *AEGIS* on DoorKey-16x16 over ten seeds, averaged over the final 5% of training. Table 3 reports the attained value against its bound for each quantity, with the spread across seeds.

Quantity	Attained	Bound	Holds
η_D vs. $\ln 2$	0.0128 ± 0.0015	$\ln 2 \approx 0.6931$	Yes
η_I vs. $\mathbb{H}(a \mid t)$	0.0384 ± 0.0080	0.9530 ± 0.0439	Yes
$\mathcal{L}^{FDM} \to 0$	0.0076 ± 0.0004 (early: 0.3041)	$< 5\%$ of early value (0.0152)	Yes

Table 3: Attained values of Assumption 1's three quantities against their bounds, on DoorKey-16x16, mean $\pm$ std over ten seeds, averaged over the final 5% of training.

For *AEGIS* on DoorKey-16x16, η_D clears its bound comfortably, at roughly 1.8% of $\ln 2$. η_I attains 0.0384 against a conditional action entropy of 0.9530, well within the bound of Assumption 1.

For $\mathcal{L}^{FDM} \to 0$, the assumption states an idealized limit rather than a strict inequality; we take the error to have effectively vanished if it falls to less than 5% of its value at the start of training, rather than asserting exact equality to zero, and report the early-training value alongside the final one so the reduction is visible directly. For *AEGIS* on DoorKey-16x16, $\mathcal{L}^{FDM}$ falls from 0.3041 at the start of training to 0.0076 in the final window, a 40-fold reduction, well below the threshold of 5% of the initial value.

Proposition 1 makes the stronger, exact claim that $r_t^{local} = 0$ whenever step t revisits an already-seen observation. Direct inspection of logged transitions confirms this: on every step at which the current observation embedding coincides exactly with an entry in the episodic history, the minimizing term of Eq. (7) has a zero numerator and r_t^{local} evaluates to exactly zero, as the proposition requires. We report this as an inspection rather than as an aggregate statistic, because the mean of r^{local} over all steps is not a valid proxy for the claim: it conflates revisit and non-revisit steps, and is additionally confounded by episode length, which changes substantially over training. A step early in a freshly reset episode has a small episodic history to minimize over, and as episodes shorten over training, such steps make up a growing share of all logged steps, independent of whether r^{local} behaves as Proposition 1 predicts on any individual revisit.

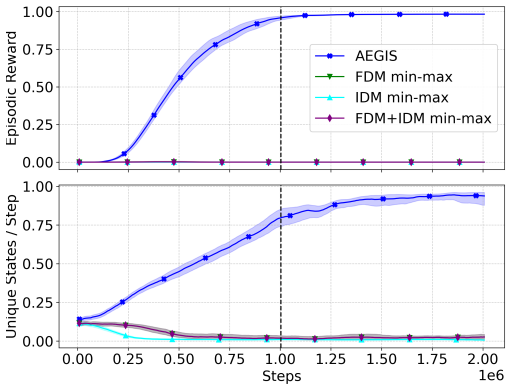


Figure 8: Ablation on alternative objectives.

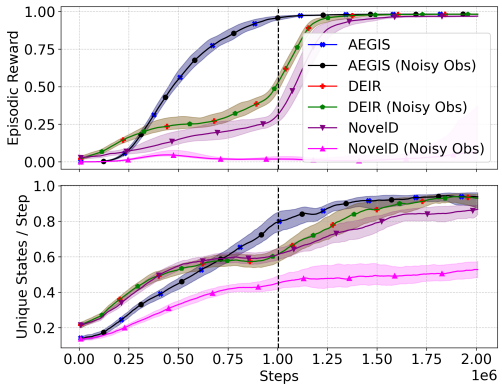


Figure 9: Comparison on noisy observations.

B Extra Ablation Studies

In this section, we list our ablation studies in greater detail. We include ablations in the intrinsic reward coefficient, but also extra combinations for the loss components in the LMDP model.

B.1 Intrinsic reward coefficient

We evaluate the impact of the intrinsic reward magnitude by scaling its contribution. Figure 10 shows that very small values make the intrinsic reward have a minuscule impact on learning, while big values, essentially matching the extrinsic reward in magnitude, cause the return to degrade rather than merely to converge slowly: at $\alpha = 1$ the agent reaches roughly 0.4 by the end of pretraining and then declines to near zero over the training phase. The effect is phase-dependent: larger coefficients accelerate learning during pretraining, where the intrinsic reward is the only signal, but destabilize the training phase by competing with the extrinsic reward. This suggests that a single coefficient for both phases is a compromise, and that annealing α at the phase boundary may recover the pretraining benefit without the later cost. We leave the investigation of using different coefficients and system components in pretraining and training for potential future work.

B.2 Optimization objective and noisy observations

Figure 8 compares the approximate bilevel Stackelberg formulation against min-max (zero-sum) alternatives on DoorKey-16x16. All three variants fail to learn. For the FDM variant, the mechanism is the one Lemma 1 identifies: the zero-sum reduction leaves the representation with the forward term alone, which a collapsed encoder minimizes. What is visible across all three is that per-step novelty *declines* over training rather than failing to rise, so the zero-sum objective actively degrades exploration rather than merely withholding a signal. This is the empirical counterpart of the argument in Section A: the general-sum formulation is not a stylistic preference but the reason the representation does not collapse. App. B.5 probes the latent representation directly and finds that all non-collapsed variants encode task structure linearly, while the zero-sum variant does not.

Exploration methods driven by prediction error are susceptible to the noisy-TV problem. Figure 9 repeats the comparison with Gaussian noise $N(\mu = 0, \sigma = 1)$ added to the already normalized observations, so that every observation looks novel regardless of any exploration effort. We first ran this on DoorKey-8x8 with *AEGIS* alone, whose noisy and clean curves were nearly identical; we then repeated it on the harder DoorKey-16x16 environment and added DEIR and NovelD for a full comparison. *AEGIS* is unaffected there too, its noisy and clean curves lying within each other's interquartile ranges on both return and per-step novelty. DEIR is likewise unaffected, while NovelD degrades severely. The asymmetry follows the form of the reward rather than the presence of an intrinsic bonus: for *AEGIS*, Proposition 1 gives that an agent which stops reaching genuinely new observations drives r^{local}, and with it the product r^{int}, to zero, so a persistently high prediction error cannot by itself sustain reward. NovelD's novelty difference is taken over an RND error with no such episodic factor, so noise sustains it indefinitely.

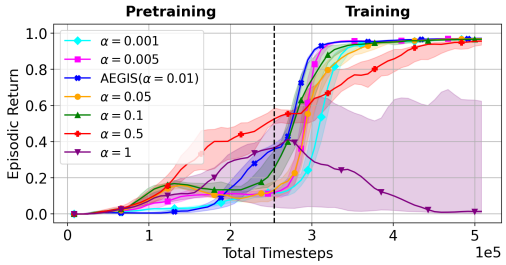


Figure 10: Ablation on r^{int} coefficient.

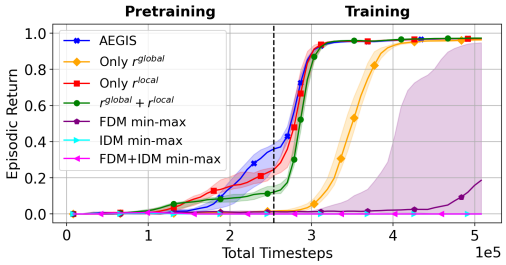


Figure 11: Ablation on r^{int} components.

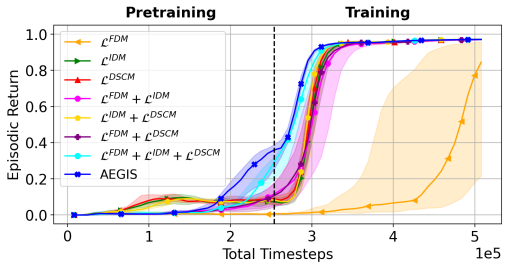


Figure 12: Ablation on $\mathcal{L}^{LMDP}$ components.

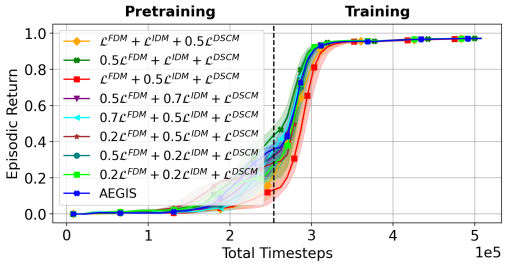


Figure 13: More fine-grained ablation on $\mathcal{L}^{LMDP}$ components.

B.3 Reward components (DoorKey-8x8)

Figure 11 repeats, on DoorKey-8x8, the reward-component ablation reported on DoorKey-16x16 in Section 5 (Figure 6). As noted there, DoorKey-8x8 separates configurations that fail outright but not those that merely converge more slowly, so we report here only what survives that limitation and defer the ordering claims to the 16x16 result.

Removing r^{local} again degrades performance, consistent with Cor. 2. All three min-max variants again fail to learn; as on DoorKey-16x16, per-step novelty declines rather than failing to rise, which is the mechanism Lemma 1 identifies for the FDM variant and is visible here across all three, together with a markedly wider dispersion across seeds than on the healthy configurations. For the reward components and the additive variant, DoorKey-8x8 does not resolve the ordering: the ablated and additive configurations converge within a narrow band of each other, and the sharper separation reported in Section 5 is specific to DoorKey-16x16.

B.4 Latent dynamics model loss components (DoorKey-8x8)

We test the effects of the loss components in isolation and in combination on DoorKey-8x8. Section 5 notes that DoorKey-8x8 separates configurations that fail outright but not those that merely converge more slowly, so this sweep is best read as consistent with the DoorKey-16x16 result in Figure 7 rather than as independent confirmation. Figure 12 shows all three components mattering in that sense: the loss-only variant reduced to $\mathcal{L}^{FDM}$ degrades severely, matching Lemma 1, while the remaining configurations converge within a comparable band. We try further combinations of component weights in Figure 13.

The performance difference across weight combinations is small, as long as all components are present. This is what Theorem 1 predicts: the weights do not enter the collapse comparison, so the exclusion holds for any positive choice of them.

B.5 Linear probing of the latent representation

Theorem 1 rules out a degenerate latent geometry but does not establish that the resulting distances carry task structure, the gap noted in Remark 2. We test this directly with a linear probe. Two linear heads are trained on the trajectory embedding z^{traj} to predict environment ground truth: the binary

key and door status, scored with cross-entropy, and the normalized Manhattan distances from the agent to the key, the door, and the goal, scored with mean squared error. The probe consumes a detached copy of the embedding, so no gradient reaches f_θ: it measures what the representation already contains rather than shaping it. This approach follows the work of Alain & Bengio (2017).

Each probe loss is plotted against its own *floor*, the loss attained by a predictor that ignores the embedding and outputs the base rate for the binary targets or the mean distance for the continuous ones. The floor is a property of the label statistics along a given policy's trajectories, not of the encoder, so it is computed separately per configuration. A loss below the floor indicates the embedding carries information about that quantity beyond what the base rate alone provides.

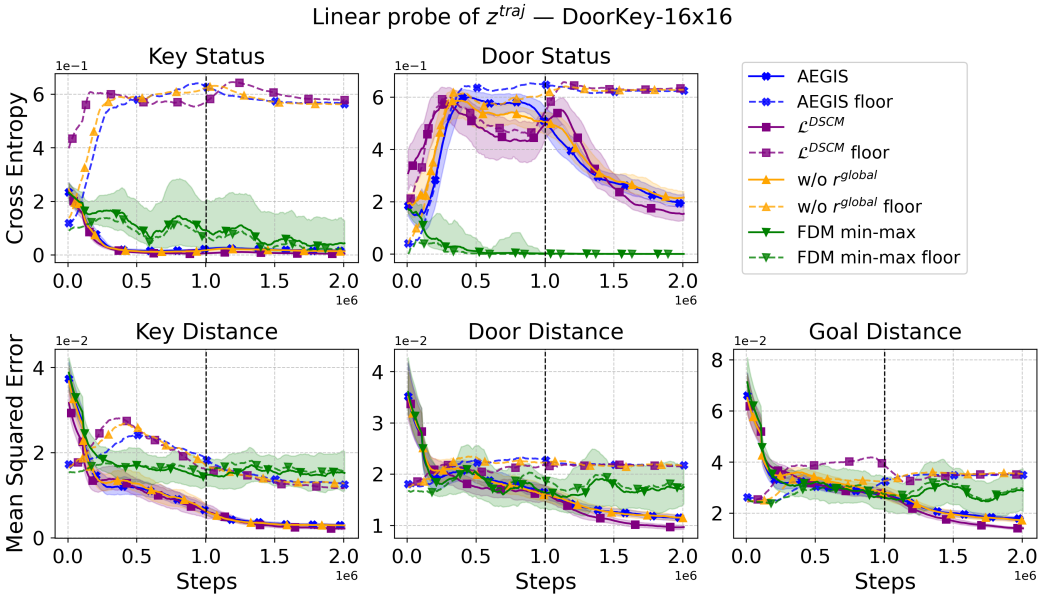


Figure 14: Linear Probing ablation to confirm representation encoding capabilities.

Figure 14 reports four configurations on DoorKey-16x16. Two results follow. First, the FDM min-max variant confirms Lemma 1 at the representation level: on all three distance targets its loss remains at or above its floor for the whole run, while the other three configurations fall well below theirs. Its two status panels are uninformative and we do not read anything from them, since a policy that rarely opens the door drives the base rate toward zero and with it the entropy floor, making proximity to the floor trivially attainable. The distance floors are immune to this, as they inherit their variance from the per-episode randomization of object positions regardless of agent behavior. Second, all three non-collapsed configurations encode key status and all three distances well below their floors, so the composite objective does produce a geometry from which task structure is linearly decodable. Goal distance separates from its floor only after the phase boundary, which is consistent with the goal acquiring significance only once the extrinsic reward is available.

We draw no conclusion about representation quality *among* the three non-collapsed configurations. Their losses are interleaved and within one another's interquartile ranges on every target, and the $\mathcal{L}^{DSCM}$-only variant decodes at least as well as the full objective on all three distances despite being metric-blind by Lemma 2. Metric sensitivity concerns invariance under affine reparametrization rather than linear decodability, so the two are not in conflict, but the probe supplies no evidence that $\mathcal{L}^{FDM}$ improves the decodability of the latent geometry. The positive alignment result discussed in Remark 2 accordingly remains open.

C Detailed Experimental Results

For the detailed experiments, we design four experimental cases by allocating different portions of the training budget for pretraining and training as:

- No pretraining, training only (NPT),
- One quarter pretraining and three quarters training (1QPT),
- Half pretraining, half training (HPT),
- Three quarters pretraining one quarter training (3QPT)

We again compare *AEGIS* against five exploration-driven methods: RND, ICM, NGU, DEIR, NovelD, along with the base PPO algorithm.

In Figure 15 we show the results from our extended experiments in Minigrid, and in Figure 16 the results from our extended experiments in Procgen.

D Experiment info & Reproducibility

This section describes the environments tested in greater detail and provides the information necessary to replicate our experiments.

D.1 Computing Environment

The experiments were conducted in a desktop computer with CPU: Intel i9-14900KF, RAM: 64GB, GPU:NVIDIA GeForce RTX 4090 (24 GB memory). Training the models for this computer has the following approximated time costs to gather one batch of the rollout data and train the model until updated once. The results are based on Minigrid tasks.

	PPO	ICM	RND	NGU	NovelD	DEIR	AEGIS
Rollout Time (sec)	1.2	1.8	1.7	4.3	2.4	3.7	3.9
Train Time (sec)	0.044	0.09	0.06	0.1	0.06	0.12	0.13

Table 4: Rollout and training times per update step for the baselines we use.

D.2 Details on Environments

In the following paragraphs, we describe in more detail the environment in which we conduct our tests. You can refer to Figure 17 for visual demonstrations.

The **BlockedUnlockPickup** environment has two 6x6 rooms of equal dimensions with a locked door placed randomly along their connecting wall. A ball is placed in front of the door, functioning as an obstacle for the agent. The agent is randomly placed in the room on the left, along with a key that opens the blocked locked door. The agent must move the obstacle from the door, retrieve the key, unlock the door, drop the key, and then pick up the box from the other room. The state space of this environment is relatively small, but accessing all of it requires successful completion of subtasks in a particular order and memory capabilities to remember what tasks have already been completed. The difficulty of this environment comes from the number of successive tasks the agent needs to perform in order to access the second room and go to the exit, while always being aware of the completed tasks. Random exploration makes it highly unlikely to complete all the sub-tasks in the correct order and successfully reach the second room that easily leads to the goal.

In the **DoorKey-16x16** environment, the agent is placed in a random place in a square room with dimensions 16x16. The room is separated into two smaller chambers, with the agent on the left one and the exit on the right. Between the two chambers is a wall with a locked door, the key of which is placed randomly in the same room as the agent. The goal is to pick up the key, unlock the door, and then reach the exit. This task is simpler than the BlockedUnlockPickup, as the door isn't blocked, and the goal is reaching the exit instead of keeping track of the possession of the key that needs to be dropped before the box is picked up. The longer training time is due to room sizes, which make this version much sparser than the BlockedUnlockPickup environment.

The **KeyCorridor-S6R3** environment has a long corridor in the middle with one room with 6x6 dimensions on each side, repeated by three rows for a total of six rooms. The rooms are connected with the corridor, and each other by locked doors of different colors, and the keys–which also come

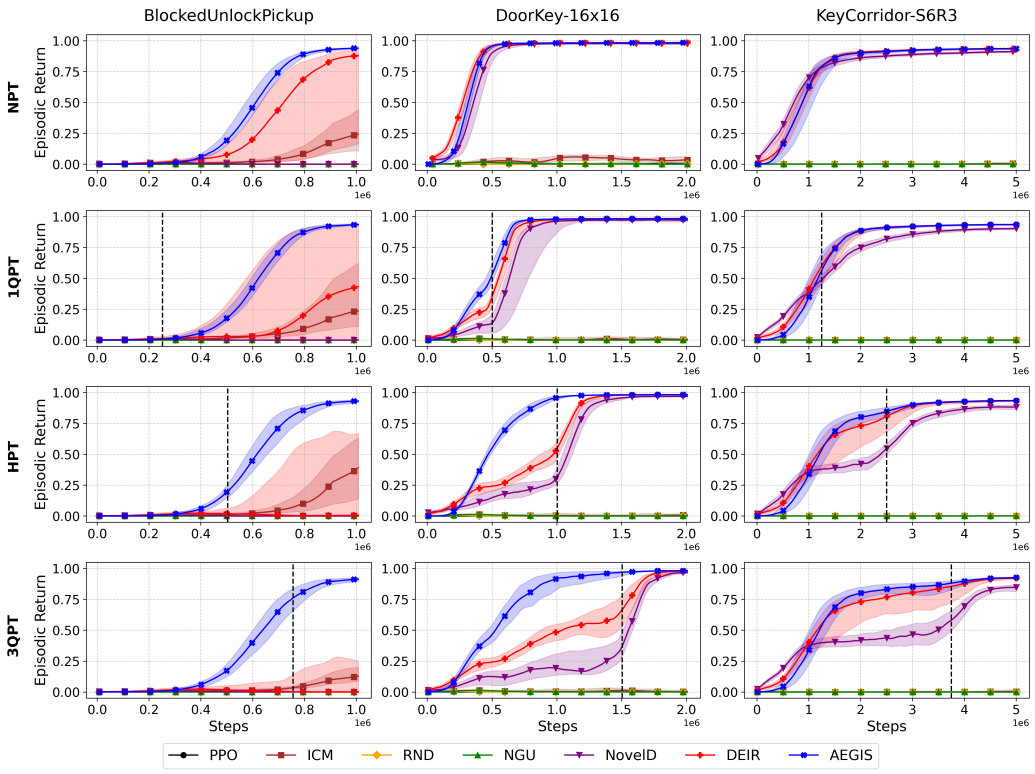


Figure 15: Detailed experiment results in Minigrid.

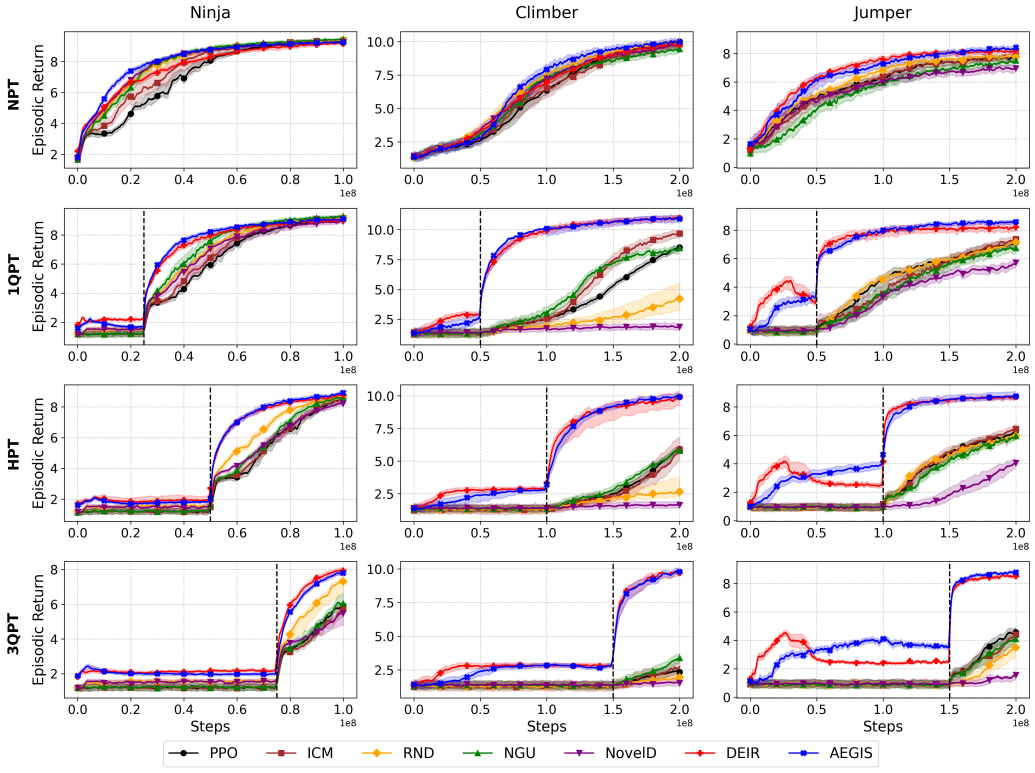
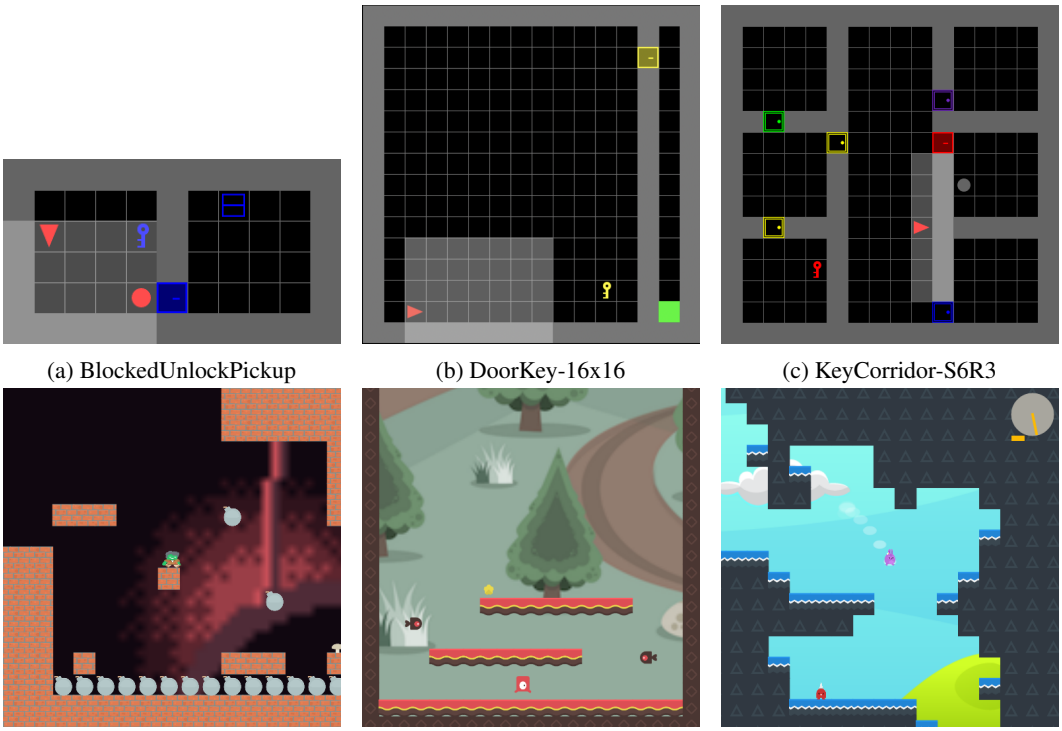


Figure 16: Detailed experiment results in Procgen.



(a) BlockedUnlockPickup (b) DoorKey-16x16 (c) KeyCorridor-S6R3

(d) Ninja (e) Climber (f) Jumper

Figure 17: Experimental environments: MiniGrid (top row) focus on logic and memory in grid-worlds, while Procgen (bottom row) environments test visual generalization in high-dimensional pixel spaces.

into different colors– are randomly placed around the environment. The agent is placed in the corridor and needs to navigate the environment to find the correct key that opens the door to the exit. This is the most difficult environment we test, and one of the most difficult environments in Minigrid.

The **Ninja** environment is a fast-paced platformer where the agent must navigate a series of procedural obstacles to reach an exit. Unlike standard platformers, the agent's movement includes high-jump capabilities and a "dash" mechanic to clear large gaps or avoid hazards. The environment is populated by narrow platforms, walls, and stationary or moving enemies. Success requires precise temporal control and spatial awareness, as falling off platforms usually results in immediate episode termination. The difficulty arises from the procedural generation of gap widths and platform heights, which prevents the agent from memorizing a specific path and instead forces it to learn a generalized policy for jump-timing and hazard avoidance.

In the **Climber** environment, the agent must navigate vertically through a series of platforms to collect stars while avoiding lethal obstacles like spikes and hostile creatures. The layout is structured as a vertical tower where the agent must constantly move upward. This environment tests the agent's ability to plan a trajectory in a 2D space where gravity is a constant factor. The primary challenge lies in the sparse reward structure—stars are often placed in risky positions—and the "one-way" nature of vertical progression; missing a platform can lead to a significant loss of progress or death. It requires the agent to prioritize long-term survival over immediate reward collection.

The **Jumper** environment features a high-contrast, platform-based layout where the agent must reach a goal located at the end of a procedurally generated level. The mechanics are centered on jumping across varied terrain while navigating "wraparound" screen edges or complex platform configurations. Compared to Ninja, the movement is often more floaty, and the level design focuses more on navigation through a maze-like structure of platforms rather than high-speed combat. The complexity stems from the visual diversity and the need for the agent to identify "safe" landing zones amidst background noise and moving hazards, testing the robustness of the agent's visual representations and its ability to handle varied jump physics.

D.3 PRETRAINING

In Table 6 we denote the hyperparameters of *AEGIS* and the other tested methods. For easier comparisons between the different algorithms we fixed as many hyperparameters as we could even across different environments. Although this doesn't lead to the best possible performance for any particular environment, it enables easier comparisons between algorithms and across the increasing environment difficulty making easier to attribute the observed differences to the algorithmic logic. In Figure 18 we denote the architecture for the encoder for the Minigrid environment, actor, critic, inverse dynamics model, forward model, and discriminator contrastive model. For the Procgen environment we use the IMPALA CNN architecture as displayed in Figure 19.

The algorithmic logic of *AEGIS* is presented in Algorithm 1 below:

Algorithm 1 AEGIS

Input: updates N, rollout length T, intrinsic reward coef. α
Initialize: LMDP params. $\theta = (\theta_{enc}, \delta, \zeta, \omega)$, policy params. ϕ, value params. ψ
$Q \leftarrow \emptyset$ {Novel-observation queue}
for $k \in \{0, 1, \ldots, N-1\}$ **do**
  $\mathcal{D} \leftarrow \emptyset$ {Reset rollout buffer}
  for $t \in \{0, 1, \ldots, T-1\}$ **do**
    if o_t starts a new episode **then**
      $E^{obs} \leftarrow \emptyset, \quad E^{traj} \leftarrow \emptyset$ {Reset episodic histories}
    end if
    $a_t \sim \pi_\phi(f_\phi(o_t))$ {Sample action from current policy}
    $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t), \quad o_{t+1} \sim q(\cdot \mid s_{t+1})$ {Environment step}
    $r_t \leftarrow \mathcal{R}(s_t, a_t, s_{t+1})$ {Extrinsic reward; 0 during pretraining}
    $r_t^{global} \leftarrow r^{global}(h_\delta, f_\theta, o_t, a_t, o_{t+1})$ {Eq. (9)}
    $r_t^{local} \leftarrow r^{local}(f_\theta, E^{obs}, E^{traj}, o_{t+1})$ {Eq. (7)}
    $r_t^{int} \leftarrow \alpha \cdot r_t^{local} \cdot r_t^{global}$ {Eq. (5)}
    $r_t \leftarrow r_t + r_t^{int}$
    if $r_t^{int} > \langle r^{int} \rangle_t$ **then** {Novelty vs. running average}
      $Q \leftarrow Q \cup \{o_t\}$ {Add novel observation; oldest dropped when full}
    end if
    $E^{obs} \leftarrow E^{obs} \cup \{z_t^{obs}\}, \quad E^{traj} \leftarrow E^{traj} \cup \{z_t^{traj}\}$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(o_t, a_t, o_{t+1}, r_t)\}$
  end for
  Update LMDP w.r.t. $\mathcal{L}^{FDM}_{h_\delta, f_\theta}, \mathcal{L}^{IDM}_{q_\zeta, f_\theta}, \mathcal{L}^{DSCM}_{g_\omega, f_\theta}$ using $\mathcal{D}, Q$ {Eq. (1)}
  Update PPO w.r.t. $\mathcal{J}^{CLIP}_{\pi_\phi, f_\phi}, \mathcal{L}^{PPO}_{V_\psi, f_\phi}, \mathcal{H}^{PPO}_{\phi}$ using $\mathcal{D}$
end for

D.4 TRAINING

In RFPT, the policy is trained to maximize an intrinsic reward function, which in this case is driven by exploration and representation error. With the intrinsic reward being exploration-focused, we chose Proximal Policy Optimization (PPO), an on-policy algorithm for lighter memory usage (Algorithm 2). PPO's loss function entropy term, although limited for sophisticated exploration, ensures sufficient policy variance (Schulman et al., 2017). The policy maintains its own encoder f_ϕ, architecturally identical to the LMDP's f_θ but updated only by the PPO objective; the two interact through the intrinsic reward and the data the policy collects, not through shared parameters. *AEGIS* combines informative state representations with efficient exploration-related features, expanding the capabilities of PPO to more complex and exploration-heavy environments. Notably, *AEGIS* can be paired with other RL algorithms. We restate the algorithm for the PPO algorithm we use for the training phase in our experiments (Algorithm 2), along with the hyperparameters used across all our experiments (Table 5).

Algorithm 2 PPO

1: Input: initial policy parameters ϕ_0, initial value function parameters ψ_0
2: **for** $k = 0, 1, 2, ...$ **do**
3: Collect a set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi_{\phi_k}$ in the environment.
4: Compute rewards-to-go $\hat{R}_t$.
5: Compute advantage estimate $\hat{A}_t$ (using any method of advantage estimation) based on the current value function V_{ψ_k}.
6: Update the policy by maximizing the PPO-Clip Objective:

$$\phi_{k+1} = \arg\max_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\Big(\frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_k}(a_t|s_t)} A^{\pi_{\phi_k}}(s_t, a_t),\ g\big(\epsilon, A^{\pi_{\phi_k}}(s_t, a_t)\big)\Big),$$

typically via stochastic ascent with Adam
7: Fit value function by regression on mean-squared error:

$$\psi_{k+1} = \arg\min_{\psi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \Big(V_{\psi_k}(s_t) - \hat{R}_t \Big)^2$$

typically via some gradient descent algorithm
8: **end for**

Parameter	Values
Encoder Module	3 conv, 1 FC, 1 GRU
Conv. channels	[32, 64, 64]
Conv. kernels	[2, 2, 2]
Conv. stride	[1, 1, 1]
Conv. padding	[0, 0, 0]
Latent Dim.	64 (via FC)
Normalization	Batch Norm
Activations	ReLU, Softmax
GRU size	64
Policy MLP	[128, $\|A\|$]
Value MLP	[128, 1]
Rollout size	512
Workers	16
Learning iter.	1
Disc. rate γ	0.99
λ_{GAE}	0.95
Adv. norm	yes (0.9 mom.)
Minibatch	512
Update epochs	4
CLIP coef.	0.2
Entropy coef.	1e-2
VF coef.	0.5
Target KL	No
Adam LR	$3e-4$
Adam ϵ	$1e-5$
Ext. Rew. Coef	0 (pre) / 1 (train)

Table 5: PPO hyper-parameters.

Parameter	Values
AEGIS	
IR coef. α	$1e-2$
IR norm	yes (0.9 mom.)
Obs queue	$1e5$
IDM loss w_I	0.5
FDM loss w_F	0.5
DSCM loss w_D	1.0
Global diff	yes
DEIR	
IR coef. α	$1e-2$
IR norm	yes (0.9 mom.)
Obs queue	$1e5$
NovelD	
IR coef. α	$3e-2$
IR norm	yes (0.9 mom.)
RND err norm	no (total avg)
NGU	
IR coef. α	$1e-3$
IR norm	yes (0.0 mom.)
RND lifelong	yes
RND err norm	yes (total avg)
Momentum	0.997
RND	
IR coef. α	$3e-3$
IR norm	yes (0.9 mom.)
RND err norm	no (total avg)
ICM	
IR coef. α	$3e-3$
IR norm	yes (0.9 mom.)
Forward loss	0.2

Table 6: Method-specific parameters.

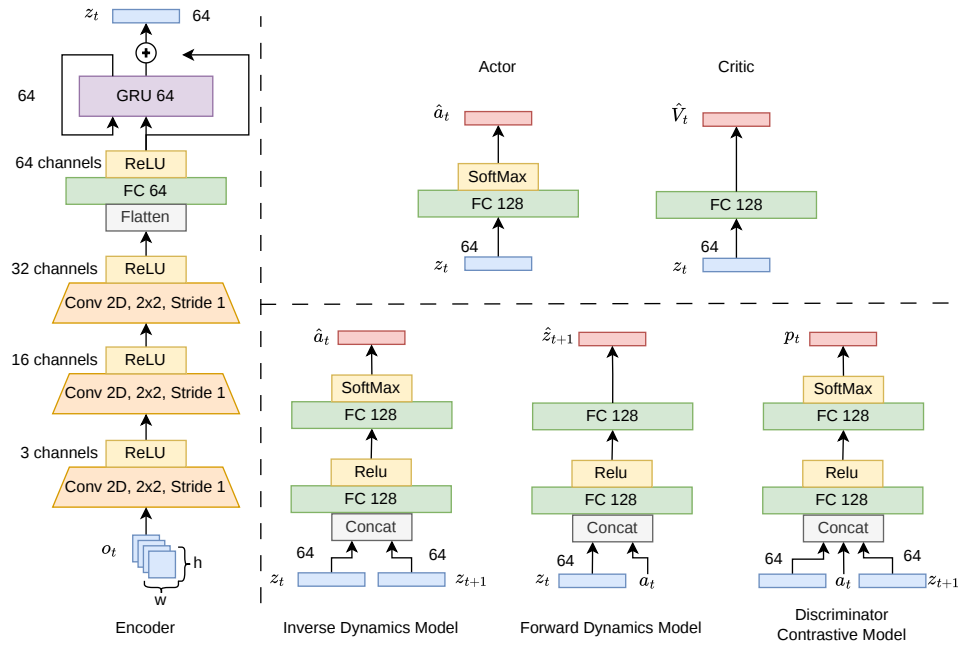


Figure 18: Architectures of the Encoder Module, Actor, Critic, Inverse Dynamics Model (IDM), Forward Dynamics Model (FDM) and Discriminator Contrastive Model (DSCM) for the Minigrid environments.

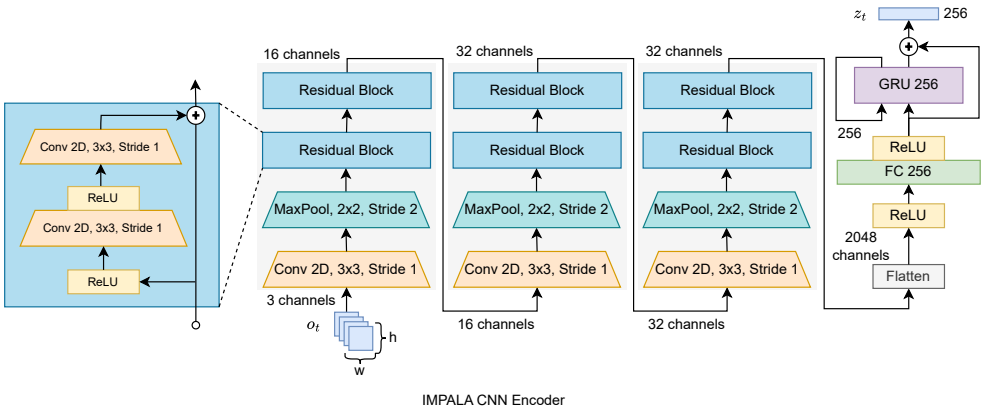


Figure 19: Architecture of the IMPALA Encoder Module for the Procgen environments.