DIFFERENT INTRINSIC DRIVES, SIMILAR SKILLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Different intrinsic drives can lead agents to learn similar skills without rewards specified for each skill. We develop HoDA (Homeostatically Driven Agent), a world-model-based agent, and compare the capabilities that agents develop with different intrinsic rewards in Crafter and two Atari games. We use achievement evaluation and behavioral statistics to examine which skills emerge across methods and how often they occur. In Crafter, intrinsic drives based on homeostasis, novelty, and predictability all support similar skills in physiological maintenance, tool use, and terrain modification. HoDA is competitive in survival and tool use and also explores broadly without novelty rewards. This study provides systematic behavioral evidence that different intrinsic drives can produce similar skills, offering a basis for understanding the relationships among intrinsic drives and further developing these shared capabilities.

1 INTRODUCTION

Autonomous agents need to learn a range of useful behaviors through interaction with their environment, but task rewards are usually designed around predefined goals, and learning mainly serves those goals (Sekar et al., 2020). In open-ended environments, the full range of needed behaviors is difficult to specify in advance, making hand-designed rewards for each environment hard to scale (Burda et al., 2019a). Intrinsic drives offer an alternative by supporting autonomous exploration and behavior learning before a specific task is given (Liu & Abbeel, 2021). We use *intrinsic drives* to refer to drives whose rewards arise from the agent's own state, interaction experience, or learning process.

Existing intrinsic drives use different criteria to assign rewards to experience. The Intrinsic Curiosity Module (ICM) (Pathak et al., 2017) and Random Network Distillation (RND) (Burda et al., 2019b) use prediction error as an intrinsic reward to encourage exploration. Active Pre-Training (APT) (Liu & Abbeel, 2021) encourages high entropy in the experienced state distribution, Exploration via Elliptical Episodic Bonuses (E3B) (Henaff et al., 2022) emphasizes within-episode diversity, and Plan2Explore (P2E) (Sekar et al., 2020) rewards ensemble disagreement in future-state predictions. Other intrinsic drives do not directly reward novelty. Empowerment (Klyubin et al., 2005) measures control over future sensory outcomes, whereas Surprise Minimizing Reinforcement Learning (SMiRL) (Berseth et al., 2021) encourages predictable experience. These methods use different forms of intrinsic reward and assign value to different experiences.

Another class of drives arises from an organism's basic need to sustain itself. Organisms maintain relatively stable internal physiological states in changing environments through *homeostasis*. Animals regulate bodily state through behaviors such as feeding and drinking, and the value of the same resource changes with their current needs (Keramati & Gutkin, 2014). Physiological needs thus provide an objective for autonomous learning by specifying the bodily states to maintain and letting the agent learn how to meet these needs. Homeostatic reinforcement learning turns this idea into a computable reward. Keramati & Gutkin (2014) define drive from deviations of physiological states from setpoints and reward as drive reduction. Yoshida & Kuniyoshi (2025) apply model-free homeostatic reinforcement learning to Crafter. Homeostasis thus provides a basis for intrinsic reward distinct from novelty, prediction error, control, and predictability.

Although these intrinsic drives differ in their reward definitions, prior work shows that the agents can learn similar behaviors. In the same energy-constrained environment, both the maximum-occupancy principle and empowerment lead agents toward food when energy is low

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.

(Ramírez-Ruiz et al., 2024). RND and P2E in Crafter also exhibit drinking, rest, and wooden-tool crafting (Hafner, 2022). Homeostatic drive can support combat and shelter construction beyond immediate replenishment (Yoshida & Kuniyoshi, 2025). These findings motivate us to ask whether similar skills arise across different intrinsic drives and extend to tool use and terrain modification that require several consecutive steps.

We use *similar skills* to mean comparable behaviors learned by agents trained with different intrinsic rewards, such as crafting a tool and then using it in combat. Their frequency, choice of tool, and execution can still differ. Identifying these shared skills can help us understand how they develop and how they can support more complex behavior.

Meeting physiological needs requires accounting for the later consequences of actions. Seeking supplies may consume resources before yielding benefits. By predicting future states and rewards, world models offer a way to learn these cumulative effects. P2E demonstrates their use for task-reward-free exploration (Sekar et al., 2020). Using DreamerV3 (Hafner et al., 2025) as a common backbone, we develop HoDA (Homeostatically Driven Agent). HoDA combines multidimensional physiological deficits into a homeostatic drive, uses its reduction as reward, and learns behavior with a world model. In Crafter, we compare HoDA with five other intrinsic methods and a task-reward reference, examining shared skills and their differences through achievements, survival, tool use, physiological maintenance, sleeping in shelters, and spatial exploration. We also examine whether different intrinsic drives produce similar functional behaviors in the Atari games Seaquest and Ms. Pac-Man.

In Crafter, agents trained with all six intrinsic methods acquire supplies, rest, fight with self-crafted weapons, and enclose areas in which they sleep. Their shared skills include behaviors measured by the built-in achievements as well as modifying and using their surroundings. Similar physiological rhythms also appear under different intrinsic drives, with food levels showing similar cycles across methods. HoDA is competitive in survival and tool use and also explores broadly without an explicit novelty reward. Together, skill performance, physiological rhythms, and spatial exploration reveal behavioral connections among different intrinsic drives.

Our contributions are:

1. We develop HoDA, a world-model-based, homeostatically driven agent that uses physiological needs to drive autonomous learning of a range of useful behaviors.
2. We systematically compare the behavioral capabilities learned with multiple intrinsic methods and quantify tool use and terrain modification beyond native achievements.
3. We find that different intrinsic drives support similar skills ranging from physiological maintenance to tool use and terrain modification, providing empirical evidence for studying the relationships among these drives.

2 Related Work

2.1 Intrinsic Drives and World Models

Intrinsic drives draw on different sources of value. Oudeyer & Kaplan (2007) distinguish computational approaches to intrinsic motivation based on prediction, novelty, and learning progress. Oudeyer et al. (2007) guide autonomous exploration through improvements in predictive competence. The methods introduced above emphasize novel experience, control, or predictability, providing distinct principles for comparing skills across intrinsic drives.

World models have also provided a common framework for comparing methods that use intrinsic rewards. Rajeswar et al. (2023) compare unsupervised reinforcement learning methods in a shared Dreamer framework, mainly evaluating task adaptation after pretraining. Ferrao & Cunha (2025) study world models with intrinsic motivation in Crafter and other environments, focusing on task return and transfer. We instead examine which skills agents learn in common through continued training with intrinsic rewards. Appendix A.1 reviews extensions to novelty, information gain, and learning from experience.

2.2 Homeostatic Reinforcement Learning

Homeostatic reinforcement learning links behavioral value to bodily needs. Keramati & Gutkin (2014) define drive from deviations between physiological states and their setpoints, and define reward as drive reduction, making the value of the same outcome depend on the internal state. Dulberg et al. (2023) study how integrated versus modular value learning for multiple needs affects exploration and adaptation. Yoshida et al. (2024) show that bodily dynamics and need weights shape long-term feeding behavior. These studies establish the theoretical basis for our reward design, with further work on anticipatory regulation and delayed bodily feedback discussed in Appendix A.2.

Yoshida & Kuniyoshi (2025) develop a model-free homeostatic agent in Crafter, showing that bodily needs can support replenishment, combat, and shelter construction. We develop HoDA using a world model and compare homeostatic and other intrinsic drives under a shared backbone to examine whether different drives can produce similar skills.

2.3 Evaluating Behavior and Capabilities under Intrinsic Drives

Some studies have found behavioral similarities across intrinsic drives in comparisons of specific methods and behaviors. The original Crafter study (Hafner, 2022) reports that both RND and P2E complete achievements involving replenishment, rest, and wooden-tool crafting. Ramírez-Ruiz et al. (2024) compare food-seeking under the maximum-occupancy principle and empowerment, finding that both seek food when energy is low despite different spatial behavior. Moreno-Bote et al. (2026) further discuss replenishment and avoidance of terminal states in terms of preserving future opportunities for action.

Cross-method studies also analyze exploration strategies and their dependence on the environment. Gruaz et al. (2025) compare six intrinsic rewards across different environment structures using state coverage, model learning, and visitation distributions, finding that performance and exploration strategies vary with the environment. Lidayan et al. (2025) compare humans with agents such as NovelD and APT in a modified Crafter environment using achievements and information-theoretic metrics. We include homeostatic drive in the cross-method comparison to examine whether these behavioral similarities extend to multistep behaviors such as tool use and shelter construction and use, and quantify how often these behaviors occur. Here, *skills* refer to behaviors that emerge during training, not separately selectable policies, a distinction discussed in Appendix A.2.

3 HoDA

3.1 Environment and Physiological State

We introduce HoDA's physiological state and reward construction using Crafter (Hafner, 2022), an environment with physiological needs and resource consumption. In this partially observable environment, the agent receives an image observation o_t and selects an action a_t at time t. Its food, water, energy, and health are represented by a four-dimensional physiological vector x_t. Each dimension ranges from 0 to 9, with higher values indicating greater reserves or better health. A value of 0 for food, water, or energy indicates that the corresponding reserve is depleted, while zero health means death. A value of 9 means that the corresponding variable is at its maximum. The agent observes a rendered image containing its local view and status bar. The four physiological scalars are not provided as separate inputs and are used only to compute HoDA's intrinsic reward.

3.2 Multidimensional Homeostatic Reward

Let $x_i^{\max}$ be the maximum value of physiological variable i. We normalize each physiological variable as $v_{t,i} = x_{t,i}/x_i^{\max} \in [0, 1]$. Let σ_i be the setpoint of normalized physiological variable i. The corresponding normalized physiological deficit is

$$d_{t,i} = \max(0, \sigma_i - v_{t,i}). \tag{1}$$

We aggregate squared deficits across dimensions with fixed weights w_i to obtain the homeostatic drive

$$D_t = \sum_{i=1}^{m} w_i d_{t,i}^2. \tag{2}$$

Here, m is the number of physiological dimensions. In Crafter, $m = 4$ and $x_i^{\max} = 9$ for all dimensions. In the main Crafter experiments, we set $\sigma_i = 1$ and $w_i = 1$, using the maximum value as each setpoint and weighting all normalized deficits equally. This uniform setting reduces per-dimension tuning and supports comparison with other intrinsic drives under a simple, fixed homeostatic reward configuration.

Squaring each deficit in Eq. (2) makes the immediate reward contribution of replenishment depend on the current need. Within a dimension, the same reduction in physiological deficit produces a larger reduction in drive when the initial deficit is greater, while its marginal benefit diminishes as the deficit is removed. This property is consistent with the analysis of need and replenishment value in homeostatic reinforcement learning (Keramati & Gutkin, 2014). Fig. B1 illustrates the reward construction, and Appendix B.2 analyzes its one-dimensional properties.

After action a_{t-1}, the homeostatic reward is

$$r_t^{\mathrm{HoDA}} = D_{t-1} - D_t. \tag{3}$$

Eq. (3) follows Keramati & Gutkin's (2014) definition of reward as homeostatic drive reduction. It measures the net improvement across physiological needs. Reward is positive when drive decreases, negative when it increases, and zero when it is unchanged. For example, drinking while thirsty contributes positively, but damage received at the same time may offset this benefit. The reward for an action therefore depends on the current bodily state and the actual changes after interaction, rather than a fixed value assigned in advance.

The one-step homeostatic reward measures the change in drive over a single transition, whereas the value of an action also depends on its effects on future states and rewards. HoDA therefore maximizes the expected discounted return of homeostatic rewards:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \gamma^{t-1} (D_{t-1} - D_t) \right]. \tag{4}$$

Here, T is the number of interaction steps from the start of an episode to termination, including the terminal transition. The discount factor $\gamma \in (0, 1)$ makes the same reduction in need more valuable when it occurs earlier. Appendix B.3 relates this return to cumulative discounted homeostatic drive.

3.3 Homeostatic Reward Prediction and Policy Learning

HoDA uses DreamerV3's recurrent state-space model (Hafner et al., 2025) to infer a latent state s_t from images and interaction history and predict dynamics, rewards, and continuation probabilities. The reward prediction head learns to predict homeostatic rewards computed from real interactions using Eq. (3). Its prediction loss is optimized jointly with the other world-model losses. Starting from latent states inferred from real experience, the policy selects actions and the world model predicts subsequent states to generate imagined trajectories. Predicted homeostatic rewards and bootstrapped value estimates are then used to compute returns for policy learning, applying the long-term objective in Eq. (4).

4 Experimental Setup

4.1 Environment and Compared Methods

We train and evaluate agents in the official Crafter environment (Hafner, 2022) using 64×64-pixel RGB observations and the native action space. Each world contains 64×64 map cells, with resource consumption and enemy threats. Its 22 native achievements are used to evaluate capabilities. We also compare behaviors in the Atari games Seaquest and Ms. Pac-Man (Bellemare et al., 2013), with the setup described in Appendix E.1. Environments that support multiple behaviors and their combinations allow us to compare shared skills across intrinsic drives.

Table 1: Aggregate achievements, survival, and sword crafting and use. The last two columns give the percentages of all episodes with (i) any sword crafted and (ii) a zombie defeated using a sword crafted earlier in the same episode.

Method	Crafter score	Mean achievements	Mean episode length (steps)	Crafted any sword (%)	Used crafted sword in combat (%)
HoDA	27.77 $\pm$ 1.38	13.60 $\pm$ 0.14	2232.23 $\pm$ 642.63	98.96 $\pm$ 0.60	97.79 $\pm$ 0.60
E3B	26.96 $\pm$ 4.76	13.23 $\pm$ 0.69	1188.48 $\pm$ 234.96	90.36 $\pm$ 15.00	86.07 $\pm$ 19.41
ICM	22.10 $\pm$ 0.56	10.79 $\pm$ 0.13	527.54 $\pm$ 21.80	79.82 $\pm$ 1.48	62.24 $\pm$ 5.58
RND	23.19 $\pm$ 2.65	11.67 $\pm$ 0.91	1017.34 $\pm$ 266.57	87.50 $\pm$ 3.96	78.12 $\pm$ 6.09
P2E	22.64 $\pm$ 1.84	12.59 $\pm$ 0.63	1131.51 $\pm$ 50.16	85.55 $\pm$ 4.74	75.65 $\pm$ 6.74
SMiRL	10.31 $\pm$ 3.29	8.33 $\pm$ 0.69	1090.65 $\pm$ 241.48	97.92 $\pm$ 1.19	95.57 $\pm$ 2.96
Task Reward	75.42 $\pm$ 1.60	19.64 $\pm$ 0.13	844.92 $\pm$ 126.13	99.74 $\pm$ 0.45	95.44 $\pm$ 0.60

In Crafter, we compare HoDA, E3B, ICM, RND, P2E, and an image-based implementation of SMiRL. All use a shared DreamerV3 backbone while retaining their required additional inputs and reward-learning procedures (Appendix C.1). Each intrinsic method is trained using only its own intrinsic reward. We also train the same backbone with the environment's native reward as a task-reward reference, denoted Task Reward. Its reward includes first-time achievement rewards within each episode and native health-change terms.

4.2 Training and Evaluation

Experiments are implemented in JAX and run on NVIDIA RTX 4090 GPUs. In Crafter, all methods use the same-sized DreamerV3 backbone and are trained for 50 million environment steps with five independent random seeds per method. Each final model is evaluated for one episode in each of the same 256 environment instances, referred to as *worlds*, with frozen parameters and actions sampled from the policy distribution.

Method-level metrics are first computed over the shared worlds within each independent training run, then averaged equally across runs and reported as mean $\pm$ standard deviation. Crafter score is computed from each run's success rates on the 22 achievements. Following Agarwal et al. (2021), we estimate confidence intervals for Crafter behavioral analyses using stratified resampling over independent training runs and shared evaluation worlds. Appendix C.2 gives training and evaluation details.

5 Results

5.1 Overall Performance

We use the standard Crafter score and achievement success rates defined by Hafner (2022). Tab. 1 shows that HoDA has the highest Crafter score and longest mean survival time among the evaluated intrinsic methods. E3B achieves a similar score, while SMiRL often uses self-crafted swords in combat despite its lower score. These methods all support survival and weapon use, so these capabilities are not limited to HoDA's explicit focus on bodily needs.

Unlike Task Reward, which directly rewards first-time sword crafting and enemy defeats, none of the six intrinsic methods uses these achievement rewards, yet agents trained with each method use self-crafted swords in combat. Weapons may help shorten fights and reduce injuries, which could explain why weapon crafting and use emerge under different intrinsic drives. Training curves for all methods are provided in Appendix D.1.

Supplementary experiments in two Atari games also reveal shared behaviors across methods. In Seaquest, agents trained with several intrinsic drives surface to replenish oxygen and then dive again. In Ms. Pac-Man, agents trained with several methods collect power pellets and then eat ghosts. Appendix E provides the behavioral statistics and further analysis.

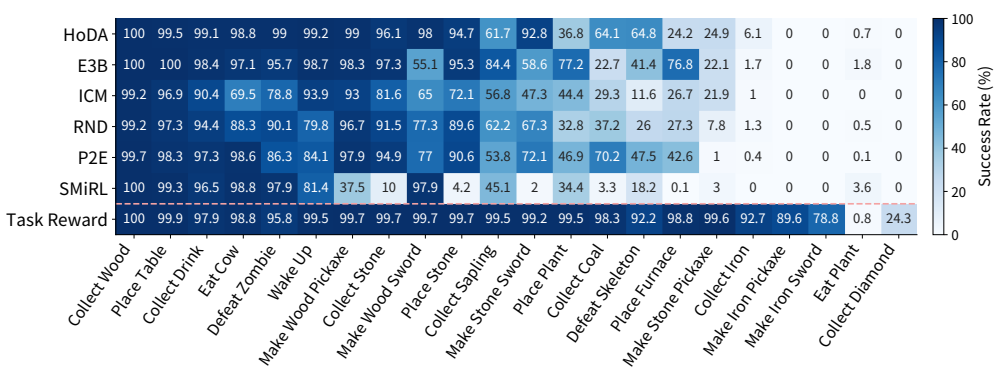

Figure 1: Success rates for the 22 native achievements (%). Each cell is the percentage of episodes that complete that achievement at least once.

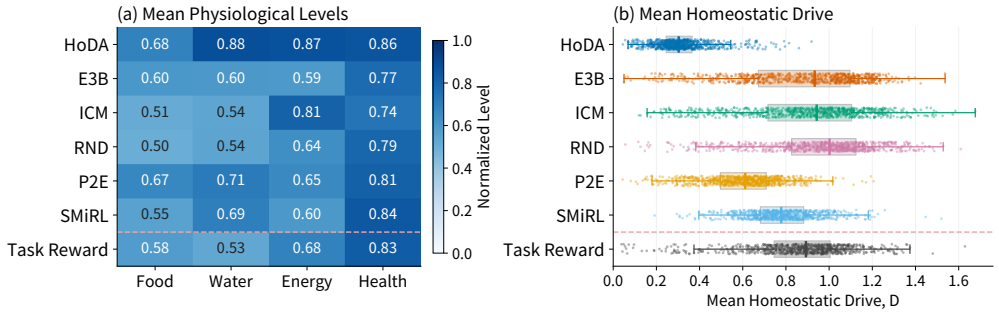

Figure 2: Physiological levels and homeostatic drive. (a) Time-averaged normalized physiological levels within each episode. (b) Mean homeostatic drive D within each episode. Each point is one episode, and boxes and center lines show the interquartile range and median. See Appendix D.2 for statistical details.

5.2 Achievement Profiles and Tool Use

Fig. 1 shows substantial overlap among the six intrinsic methods in resource gathering, replenishment, tool crafting, and combat achievements. These shared capabilities appear both with methods that encourage novel experience and with HoDA and surprise-minimizing SMiRL. The details of these skills still differ. HoDA performs better at later stages of the technology tree, such as crafting stone pickaxes and collecting iron, and more often defeats skeletons. E3B more often builds furnaces, plants crops, and collects saplings. SMiRL rarely crafts stone swords but often defeats zombies with self-crafted wooden swords. It therefore shares weapon-based combat with other methods while relying more on wooden than stone swords.

The methods also show similar limits. Several intrinsic methods occasionally collect iron, but none of the six intrinsic methods in the main comparison completes the iron-tool or diamond-collection achievements. Task Reward completes these later achievements.

5.3 Physiological Maintenance and Shelter Construction

We next compare agents' physiological states, replenishment, rest, and shelter construction across methods. Fig. 2 shows that agents trained with non-homeostatic intrinsic rewards also maintain food and health levels. HoDA maintains higher water and energy levels and has lower homeostatic drive. Appendix D.3 further shows that both HoDA and P2E are less likely to drink when the water level is at its maximum. Despite their different reward sources, both show replenishment choices associated with current bodily state.

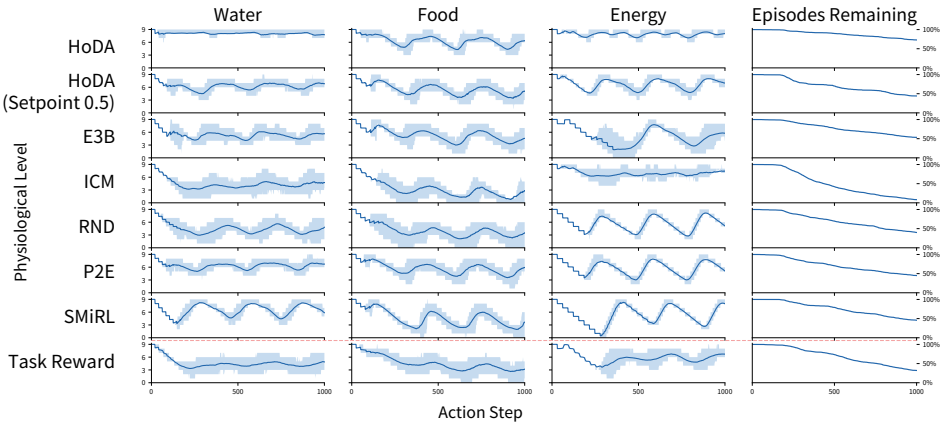

Figure 3: Physiological state over evaluation steps. The first three columns show water, food, and energy (0–9). Blue lines and light-blue bands show the mean and 25th–75th percentiles of episodes with a recorded value at each step, while the right column shows the percentage of episodes with recorded values. The second row is HoDA with setpoint 0.5. Consistent discrete depletion early in episodes makes the means stepwise, while unsmoothed percentile bands change in steps as quartiles cross integer levels. Appendix D.2 provides episode-level results and statistical details.

Fig. 3 shows that, after the initial decline, the mean food level fluctuates with a period of roughly 300 steps for every method. HoDA with setpoint 0.5, RND, P2E, and SMiRL also show similar depletion–recovery cycles in mean energy, with RND and P2E recovering at similar phases and SMiRL recovering later. HoDA's energy cycle is more pronounced at setpoint 0.5 than at the main setpoint of 1.0. These rhythms recur across independent training runs and worlds rather than arising from a single trajectory. One possible explanation is that all methods face the same environmental cycle and physiological rules. Crafter has a 300-step day–night cycle, and food availability and threats vary with light. Energy is depleted by shared physiological rules and restored through sleep. Policies may therefore learn to schedule replenishment and rest around this common cycle, regardless of their reward source. Appendix D.7 compares physiological time courses and training performance across setpoints.

Fig. 4 shows agents sleeping in self-built shelters. They place objects such as stones and crafting tables, create paths, and block entrances to form enclosed sleeping spaces. Agents trained with all six intrinsic methods enclose areas and sleep inside them. This behavior occurs in 34.8% to 93.0% of episodes across these methods and also appears under Task Reward. Similar skills across intrinsic drives thus extend beyond replenishment and rest to modifying the surroundings to create a resting space. Additional statistics show that agents trained with all six intrinsic methods return to previously used shelters to sleep again. Appendix D.4 provides these statistics and the construction sequence.

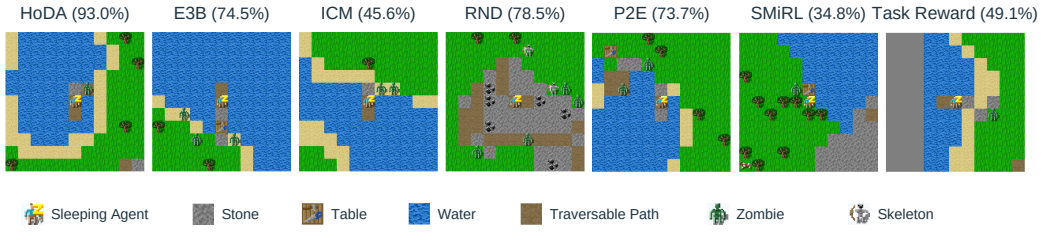

Figure 4: Examples of sleeping in shelters across methods. Parentheses give the percentage of episodes in which agents enclose an area and sleep inside it. The original nighttime scenes are dark, so uniform lighting is applied during post-processing for visibility.

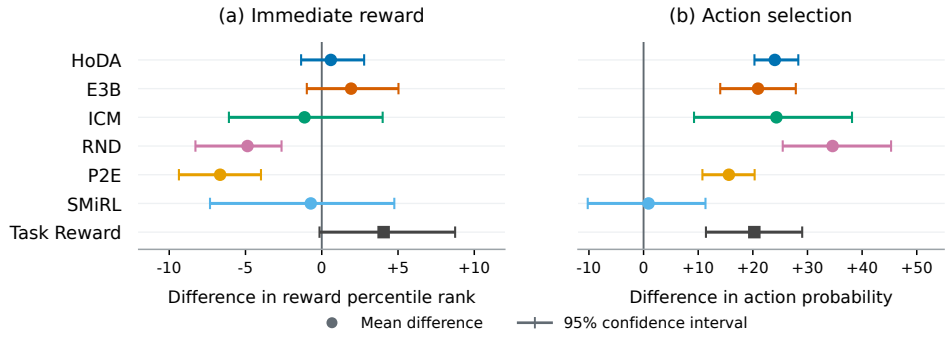

Figure 5: Shelter placements compared with matched ordinary placements. Each difference is the metric value for shelter-placement actions minus that for matched ordinary-placement actions. The left and right panels show differences in reward percentile rank and action probability, measured in percentile points and percentage points, respectively.

To better understand this shared behavior, we compare shelter-placement actions with ordinary placements of the same action type in the same world, matched for time and state. We examine the relationship between immediate rewards and policy choices. Each reward is expressed as its percentile rank among rewards during awake steps in the same episode, and action selection probabilities are given by the policy.

Fig. 5 shows that RND and P2E assign higher probabilities to shelter-placement actions despite their lower reward ranks. HoDA, E3B, ICM, and Task Reward also assign higher probabilities, but show no clear immediate-reward difference. Neither difference is clear for SMiRL. These differences may reflect how each method evaluates experience. RND rewards prediction error on observation features (Burda et al., 2019b), while P2E rewards ensemble disagreement in state-transition predictions (Sekar et al., 2020). Both signals may decrease as the models learn from relevant experience. If the observations or transitions involved in shelter construction are more familiar than those involved in ordinary placement, their immediate rewards may be lower. ICM instead measures prediction error on next-state features (Pathak et al., 2017). Agreement among models does not imply accurate predictions, so ICM need not show the same reward difference as P2E.

Although these methods assign different immediate rewards to construction actions, their agents all learn to construct shelters. Shelter construction may lead to rewards through later rest and interaction, rather than receiving a high reward at each construction step. For HoDA, rewards come from improvements in physiological state, and construction itself does not directly replenish resources. The reward decomposition in Appendix D.5 shows negative mean net rewards during construction and positive mean net rewards during subsequent sleep, consistent with this explanation. Appendix D.5 also provides the matching procedure and further analysis of each drive.

The shared use of shelters may arise from the need for safety during sleep. Under Crafter's physiology and sleep rules, sleep restores energy and health but temporarily prevents movement and attack, leaving the agent more exposed to zombies. Using terrain and placed objects to block enemies may reduce damage and interruptions during recovery. For SMiRL, fewer external disturbances may also help keep observations predictable (Berseth et al., 2021). These methods have no shared shelter reward but face the same risks while resting, which may make terrain modification valuable under different intrinsic drives.

5.4 Spatial Exploration

We next compare spatial exploration across methods, using the number of unique map cells visited per episode to measure spatial coverage. Fig. 6 shows that all methods expand beyond the starting area. HoDA visits the most cells on average over complete episodes. E3B and P2E also visit broad regions. SMiRL covers a smaller area, while ICM and RND remain more concentrated. Both HoDA and SMiRL show spatial exploration without explicit novelty or map-coverage rewards.

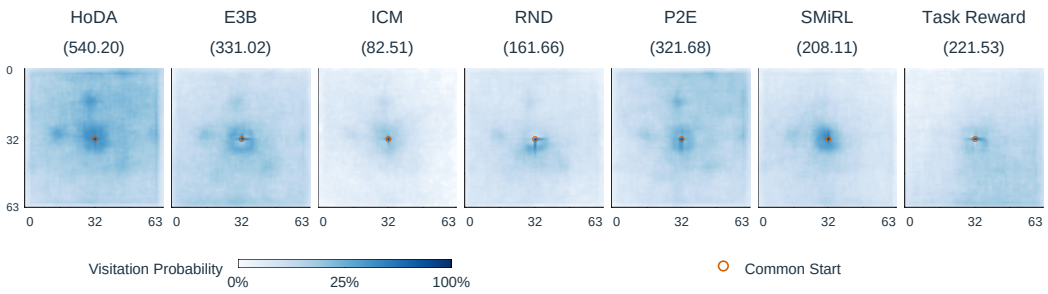

Figure 6: Spatial visitation probabilities across methods. Each cell shows the percentage of complete episodes that visited that coordinate. Parentheses give the mean number of unique cells visited per episode.

For HoDA, searching for scattered supplies and coping with resource depletion and threats may make exploration a means of meeting bodily needs. Full-episode coverage also depends on survival time and is not equivalent to exploration speed at the same action budget. Appendix D.6 provides full visitation statistics, trajectories in one shared world, and visitation and survival analyses at fixed action budgets.

6 Discussion and Limitations

6.1 Implications of Shared Skills for Intrinsic-Drive Research

The emergence of similar skills under multiple intrinsic drives shows that, in our environment and learning setup, autonomous skill learning does not depend on a single reward design. This finding suggests that future research should not only explore new intrinsic rewards but also seek to make better use of existing drives. Skills shared across drives provide a basis for studying learning methods that can make effective use of different intrinsic rewards, while differences between methods can help clarify when these learning methods apply.

6.2 From Similar Skills to More Complex Capabilities

Our study shows that intrinsic rewards alone can support the autonomous learning of multiple skills, demonstrating the potential of intrinsic drives. Yet the range of learned skills remains limited, and enabling agents to keep acquiring new skills remains an open problem. Future research should examine how skills form during training and what limits further learning.

6.3 Limitations

Our systematic behavioral comparison focuses on Crafter, with supplementary evidence from two Atari games. All experiments use the DreamerV3 learning framework. Broader comparisons are needed to determine how much the observed similarities depend on the current environments and learning backbone. HoDA also requires meaningful physiological variables that the agent can regulate through its actions. In addition, this study has not established how these skills form during training.

7 Conclusion

By developing HoDA and systematically comparing multiple intrinsic drives, we find that different drives can produce similar skills in the same environment. This finding provides empirical evidence for studying connections among intrinsic drives and motivates further study of why shared skills emerge. Intrinsic drives also show promise for autonomous skill learning without task rewards designed for each skill. Realizing this potential more fully requires identifying what prevents agents from continuing to learn new skills.

AI USE STATEMENT

Generative AI tools assisted with literature review, code implementation, and manuscript editing.

REPRODUCIBILITY STATEMENT

Training and evaluation protocols are described in Section 4.2. Appendix C provides method implementations and shared training settings, while Appendix D details the additional evaluation procedures.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research (JAIR)*, 47:253–279, 2013. doi: 10.1613/jair.3912. URL https://arxiv.org/abs/1207.4708.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/afda332245e2af431fb7b672a68b659d-Abstract.html.

Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. SMiRL: Surprise Minimizing Reinforcement Learning in Unstable Environments. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=cPZOyoDloxl.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. In *International Conference on Learning Representations (ICLR)*, 2019a. URL https://openreview.net/forum?id=rJNwDjAqYX.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by Random Network Distillation. In *International Conference on Learning Representations (ICLR)*, 2019b. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Roger Creus Castanyer, Joshua Romoff, and Glen Berseth. Improving Intrinsic Exploration by Creating Stationary Objectives. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7058bc192a37f5e5a57398887b05f6f6-Abstract-Conference.html.

Zack Dulberg, Rachit Dubey, Isabel M. Berwian, and Jonathan D. Cohen. Having multiple selves helps learning agents explore and adapt in complex changing worlds. *Proceedings of the National Academy of Sciences (PNAS)*, 120(28):e2221180120, 2023. doi: 10.1073/pnas.2221180120. URL https://doi.org/10.1073/pnas.2221180120.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity Is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Jeremias Ferrao and Rafael Fernandes Cunha. World Model Agents with Change-Based Intrinsic Motivation. In *Northern Lights Deep Learning Conference (NLDL)*, volume 265 of *Proceedings of Machine Learning Research*, pp. 66–74. PMLR, 2025. URL https://proceedings.mlr.press/v265/ferrao25a.html.

Sébastien Forestier and Pierre-Yves Oudeyer. Curiosity-Driven Development of Tool Use Precursors: a Computational Model. In *Proceedings of the Annual Conference of the Cognitive Science Society (CogSci)*, pp. 1859–1864, 2016. URL https://escholarship.org/uc/item/1z17525k.

Yao Fu, Run Peng, and Honglak Lee. Go Beyond Imagination: Maximizing Episodic Reachability with World Models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10405–10420. PMLR, 2023. URL https://proceedings.mlr.press/v202/fu23c.html.

Mana Fujiwara and Naoki Honda. Gradual proactive regulation of body state by reinforcement learning of homeostasis. *Neuroscience Research*, 223:105021, 2026. doi: 10.1016/j.neures.2026.105021. URL https://doi.org/10.1016/j.neures.2026.105021.

Lucas Gruaz, Alireza Modirshanechi, Sophia Becker, and Johanni Brea. Merits of Curiosity: A Simulation Study. *Open Mind: Discoveries in Cognitive Science*, 9:1037–1065, 2025. doi: 10.1162/opmi.a.9. URL https://push-zb.helmholtz-munich.de/frontdoor.php?la=de&source_opus=75605.

Danijar Hafner. Benchmarking the Spectrum of Agent Capabilities. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=1W0z96MFEoH.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via Elliptical Episodic Bonuses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 37631–37646. Curran Associates, Inc., 2022. doi: 10.52202/068431-2728. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f4f79698d48bdc1a6dec20583724182b-Abstract-Conference.html.

Zhibo Hou, Zhiyu An, and Wan Du. Beyond Noisy-TVs: Noise-Robust Exploration via Learning Progress Monitoring. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/803b9c4a8e4784072fdd791c54d614e2-Abstract-Conference.html.

Yuhua Jiang, Qihan Liu, Yiqin Yang, Xiaoteng Ma, Dianyu Zhong, Hao Hu, Jun Yang, Bin Liang, Bo Xu, Chongjie Zhang, and Qianchuan Zhao. Episodic Novelty Through Temporal Distance. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b111aa30ab71255946b19b6bd4e68939-Abstract-Conference.html.

Tim Joseph, Marcus Fechner, Philipp Stegmaier, Karam Daaboul, and J. Marius Zöllner. CIG: Exploration via Conditional Information Gain. arXiv preprint arXiv:2605.20878, 2026. URL https://arxiv.org/abs/2605.20878.

Isaac Kauvar, Chris Doyle, Linqi Zhou, and Nick Haber. Curious Replay for Model-based Adaptation. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16018–16048. PMLR, 2023. URL https://proceedings.mlr.press/v202/kauvar23a.html.

Aya Kayal, Eduardo Pignatelli, and Laura Toni. The impact of intrinsic rewards on exploration in Reinforcement Learning. *Neural Computing and Applications*, 37(21):16269–16303, 2025. doi: 10.1007/s00521-025-11340-0. URL https://doi.org/10.1007/s00521-025-11340-0.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811, 2014. doi: 10.7554/eLife.04811. URL https://doi.org/10.7554/eLife.04811.

Aniket Khan, Charitha Palika, and V. Srinivasa Chakravarthy. NutriRL: A Benchmark for Nutritional Regulation under Delayed State Transitions. *Reinforcement Learning Journal (RLJ)*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper78.html. RLC 2026 preproceedings.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All Else Being Equal Be Empowered. In *Advances in Artificial Life: European Conference on Artificial Life (ECAL)*, volume 3630 of *Lecture Notes in Computer Science*, pp. 744–753. Springer, 2005. doi: 10.1007/11553090_75. URL https://doi.org/10.1007/11553090_75.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: Unsupervised Reinforcement Learning Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/091d584fced301b442654dd8c23b3fc9-Abstract-round2.html.

Aly Lidayan, Yuqing Du, Eliza Kosoy, Maria Rufova, Pieter Abbeel, and Alison Gopnik. Intrinsically-Motivated Humans and Agents in Open-World Exploration. arXiv preprint arXiv:2503.23631, 2025. URL https://arxiv.org/abs/2503.23631.

Hao Liu and Pieter Abbeel. Behavior From the Void: Unsupervised Active Pre-Training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18459–18473. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/99bf3d153d4bf67d640051a1af322505-Abstract.html.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. *Journal of Artificial Intelligence Research (JAIR)*, 61:523–562, 2018. URL https://arxiv.org/abs/1709.06009.

Fryderyk Mantiuk, Hanqi Zhou, and Charley M. Wu. From Curiosity to Competence: How World Models Interact with the Dynamics of Exploration. In *Proceedings of the Annual Conference of the Cognitive Science Society (CogSci)*, volume 47, pp. 2398–2405, 2025. URL https://escholarship.org/uc/item/0f04713z.

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Alexandre Lacoste, and Sai Rajeswar. Choreographer: Learning and Adapting Skills in Imagination. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PhkWyijGi5b.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and Achieving Goals via World Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 24379–24391. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cc4af25fa9d2d5c953496579b75f6f6c-Abstract.html.

Faisal Mohamed, Catherine Ji, Benjamin Eysenbach, and Glen Berseth. Temporal Representations for Exploration: Learning Complex Exploratory Behavior without Extrinsic Rewards. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=KjYpHySlb0.

Rubén Moreno-Bote, Ralf Haefner, Jordi Galiano-Landeira, Tianming Yang, and Pedro Maldonado. How Intrinsic Motivation Underlies Embodied Open-Ended Behavior. arXiv preprint arXiv:2601.10276, 2026. URL https://arxiv.org/abs/2601.10276.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping. In *International Conference on Machine Learning (ICML)*, pp. 278–287. Morgan Kaufmann, 1999. URL https://ai.stanford.edu/~ang/papers/shaping-icml99.pdf.

Pierre-Yves Oudeyer and Frédéric Kaplan. What is Intrinsic Motivation? A Typology of Computational Approaches. *Frontiers in Neurorobotics*, 1:6, 2007. doi: 10.3389/neuro.12.006.2007. URL https://doi.org/10.3389/neuro.12.006.2007.

Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation (TEVC)*, 11 (2):265–286, 2007. doi: 10.1109/TEVC.2006.890271. URL https://doi.org/10.1109/TEVC.2006.890271.

Octavio Pappalardo. Unsupervised Learning of Efficient Exploration: Pre-Training Adaptive Policies via Self-Imposed Goals. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/85f27ca0470101e26ed732bce63ae7e9-Abstract-Conference.html.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-Supervised Exploration via Disagreement. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071. PMLR, 2019. URL https://proceedings.mlr.press/v97/pathak19a.html.

Sai Rajeswar, Pietro Mazzaglia, Tim Verbelen, Alexandre Piché, Bart Dhoedt, Aaron Courville, and Alexandre Lacoste. Mastering the Unsupervised Reinforcement Learning Benchmark from Pixels. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28598–28617. PMLR, 2023. URL https://proceedings.mlr.press/v202/rajeswar23a.html.

Jorge Ramírez-Ruiz, Dmytro Grytskyy, Chiara Mastrogiuseppe, Yamen Habib, and Rubén Moreno-Bote. Complex behavior from intrinsic motivation to occupy future action-state path space. *Nature Communications*, 15(1):6368, 2024. doi: 10.1038/s41467-024-49711-1. URL https://doi.org/10.1038/s41467-024-49711-1.

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 (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592. PMLR, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Leonardo Villalobos-Arias, Grant Forbes, Jianxun Wang, David L. Roberts, and Arnav Jhala. Minding Motivation: The Effect of Intrinsic Motivation on Agent Behaviors. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 21(1):347–356, 2025. doi: 10.1609/aiide.v21i1.36838. URL https://doi.org/10.1609/aiide.v21i1.36838.

Chengyang Ying, Huayu Chen, Xinning Zhou, Zhongkai Hao, Hang Su, and Jun Zhu. Exploratory Diffusion Model for Unsupervised Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/04b943ec09ae71cc0fac482fb3edf8d8-Abstract-Conference.html.

Naoto Yoshida and Yasuo Kuniyoshi. Unexpected Capability of Homeostasis for Open-Ended Learning. In *IEEE International Conference on Development and Learning (ICDL)*, pp. 1–8. IEEE, 2025. doi: 10.1109/ICDL63968.2025.11204447. URL https://doi.org/10.1109/ICDL63968.2025.11204447.

Naoto Yoshida, Etsushi Arikawa, Hoshinori Kanazawa, and Yasuo Kuniyoshi. Modeling long-term nutritional behaviors using deep homeostatic reinforcement learning. *PNAS Nexus*, 3(12): pgae540, 2024. doi: 10.1093/pnasnexus/pgae540. URL https://doi.org/10.1093/pnasnexus/pgae540.

Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E. Gonzalez, and Yuandong Tian. NovelD: A Simple yet Effective Exploration Criterion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 25217–25230. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d428d070622e0f4363fceae11f4a3576-Abstract.html.

Andrew Zhao, Erle Zhu, Rui Lu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Self-Referencing Agents for Unsupervised Reinforcement Learning. *Neural Networks*, 188:107448, 2025. doi: 10.1016/j.neunet.2025.107448. URL https://doi.org/10.1016/j.neunet.2025.107448.

A Related Work

A.1 Intrinsic Drives and Learning from Experience

Research on novelty and learnable information has improved exploration through several approaches to reward design. Bellemare et al. (2016) derive pseudo-counts from a density model to extend count-based exploration to high-dimensional observations. NovelD rewards changes in novelty between consecutive states to encourage agents to cross the boundary of explored regions (Zhang et al., 2021). ExDM estimates the experience distribution with a diffusion model and rewards less-explored regions (Ying et al., 2026). For predictive signals, ensemble disagreement can reduce the ambiguity of errors from a single model (Pathak et al., 2019), while LPM uses improvements in predictive competence to distinguish learnable change from unlearnable noise (Hou et al., 2026).

Episodic memory and model prediction change how novelty is assessed. ETD constructs episodic novelty from temporal distance between states (Jiang et al., 2025). GoBI uses model-predicted future states to estimate growth in the reachable set (Fu et al., 2023). CIG uses a reward that approximates conditional information gain to reduce information redundancy within a trajectory (Joseph et al., 2026). These studies mainly refine exploration signals, whereas we examine similar skills across intrinsic drives.

Other work improves how experience is organized and used. LEXA trains separate exploration and goal-achievement policies in a shared world model (Mendonca et al., 2021). ULEE improves adaptation to unseen tasks through an autonomous goal curriculum and cross-episode meta-learning (Pappalardo, 2026). Curious Replay uses curiosity signals to adjust experience replay for world-model learning (Kauvar et al., 2023). SOFE includes the statistics that determine exploration rewards in the policy state (Creus Castanyer et al., 2024), and Self-Reference supplies the policy with retrieved previously visited states (Zhao et al., 2025). These studies suggest ways to improve learning under intrinsic drives by organizing experience, changing replay, and using historical information.

A.2 Homeostatic Learning and Related Behavioral Studies

Homeostatic learning also involves time scale and delayed feedback. Fujiwara & Honda (2026) study anticipatory compensation, extinction, and relearning with a continuous homeostatic reinforcement learning model. NutriRL incorporates delayed absorption of multiple nutrients into state transitions to study credit assignment under delayed bodily feedback (Khan et al., 2026). These questions are relevant to learning longer-term physiological regulation with HoDA, but they are not the primary evidence for our cross-method skill comparison.

The environment and episode-termination settings can also change the behavior supported by the same intrinsic drive. In VizDoom, the original SMiRL study (Berseth et al., 2021) observes policies that end episodes as early as possible to return to similar initial states, and uses fixed-length episodes and disables death to prevent this behavior. In Crafter, we instead observe replenishment, rest, and sleeping in shelters. These findings suggest that understanding behavior under a predictability objective requires considering how environmental changes and termination affect future returns.

The capabilities of intrinsically driven agents have also been studied in other environments and evaluation protocols. Learning progress has been used to develop precursors of tool use in a simulated robotic arm (Forestier & Oudeyer, 2016). C-TeC uses a temporal representation reward to support task-reward-free achievement discovery in Craftax-Classic (Mohamed et al., 2026). In MiniGrid, Kayal et al. (2025) show that observation representations change the spatial coverage of agents trained with different intrinsic rewards, while Villalobos-Arias et al. (2025) study how exploration bonuses affect behavior under sparse task rewards. Mantiuk et al. (2025) compare several exploration objectives under fixed representations and Dreamer-learned representations, revealing interactions between exploration and world-model learning.

Our use of *skill* describes comparable behaviors that arise naturally under different intrinsic drives and differs from the learning target of explicit skill discovery. DIAYN learns diverse policies selected by a latent variable (Eysenbach et al., 2019). Choreographer learns skill-conditioned policies in a world model and trains a meta-controller for downstream tasks (Mazzaglia et al., 2023). URLB evaluates representations and behavioral capabilities by adaptation to downstream tasks after task-reward-free pretraining (Laskin et al., 2021). We do not learn a set of individually selectable skills

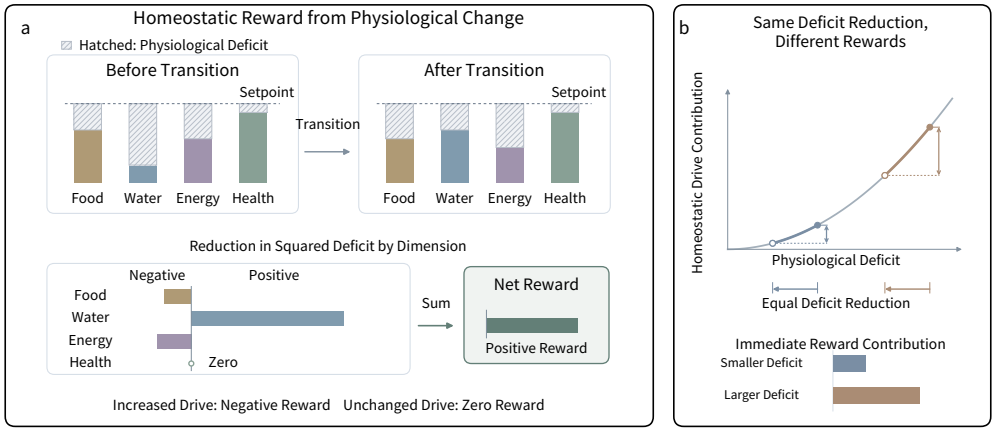

Figure B1: HoDA's homeostatic reward. (a) Each physiological dimension's contribution to net reward. (b) How the initial deficit affects the reward from recovery.

or perform downstream fine-tuning. Instead, we compare the similar behaviors that agents develop through continued training with different intrinsic rewards. Our contribution is therefore behavioral evidence across intrinsic drives, not a new skill-discovery algorithm.

B Analysis of HoDA's Reward

B.1 Reward Construction

Fig. B1 illustrates the homeostatic reward used in Section 3.2. The contribution of each dimension is the difference between its weighted squared deficit before and after an interaction.

B.2 One-Dimensional Reward Contribution

Keramati & Gutkin (2014) analyze how need intensity affects replenishment value and connect reward concavity to risk aversion. To clarify the role of our squared form, let d be the normalized physiological deficit in dimension i before recovery and let δ be its actual reduction. The drive reduction in that dimension, before minus after recovery, is

$$\Delta D_i(d, \delta) = w_i\left[d^2 - (d-\delta)^2\right] = w_i(2d\delta - \delta^2), \qquad 0 \le \delta \le d. \tag{5}$$

With the positive weights used here, a fixed recovery $\delta > 0$ produces a larger drive reduction at a larger initial deficit d. At fixed d, the marginal gain from further recovery decreases as δ grows. Thus, the immediate reward contribution of replenishment varies with current need without separately specifying rewards for every level of deprivation. Here, δ is the actual reduction in physiological deficit, not the quantity of a resource consumed. Eq. (5) describes one dimension, but recovery and depletion in other dimensions still contribute to the net reward.

Holding the initial physiological state, timing of recovery, and changes in all other dimensions fixed, the concavity of Eq. (5) in δ means that a fixed recovery amount yields at least as much immediate reward as the expected reward from a random recovery with the same mean reduction in physiological deficit. This interpretation of risk aversion concerns uncertainty in recovery outcomes, while policy choice also depends on future return.

B.3 Episode Termination and Returns

The relation between return under drive-reduction rewards and cumulative discounted drive is discussed in homeostatic reinforcement learning (Keramati & Gutkin, 2014). Relatedly, Ng et al. (1999) study conditions under which reward transformations preserve optimal policies. To make the role of episode termination explicit, consider an episode that terminates after transition T, rather

than being truncated, and define

$$G_\Delta = \sum_{t=1}^{T} \gamma^{t-1}(D_{t-1} - D_t), \qquad G_- = \sum_{t=1}^{T} \gamma^{t-1}(-D_t). \tag{6}$$

Combining adjacent terms in Eq. (6) yields

$$G_\Delta = (1-\gamma)G_- + D_0 - \gamma^T D_T. \tag{7}$$

When the initial state is fixed, D_0 is constant, but the final term $-\gamma^T D_T$ depends on termination time and terminal homeostatic drive. The relation between the two rewards therefore contains a terminal boundary term in finite episodes. Removing this term requires suitable continuing-task assumptions or conditions on the tail term. Time-limit truncation still requires accounting for the remaining value of the unfinished trajectory.

C Method Implementations and Training Settings in Crafter

C.1 Reward and Learning Path of Each Method

HoDA, E3B, and SMiRL compute rewards during real interaction and store them in replay. RND and ICM recompute rewards from replayed observations during learning. These five methods train a reward head and use its predictions to compute imagined returns. The reward head follows DreamerV3's distributional prediction and cross-entropy loss (Hafner et al., 2025). P2E directly uses ensemble disagreement as imagined reward.

RND uses the mean squared prediction error against a fixed random target network as reward. Both the target and predictor output 512-dimensional features and use three convolutional layers with 32, 64, and 64 channels, kernel sizes 8, 4, and 3, and strides 4, 2, and 1. The predictor has two additional hidden layers of 512 units each. Input images are normalized using running means and variances, with a variance floor of 10^{-4}, then clipped to $[-5, 5]$.

ICM and E3B each use four convolutional layers with 32 channels, kernel size 3, and stride 2, followed by a projection to 128 and 64 features, respectively. Their inverse dynamics networks and ICM's forward network each have one hidden layer of 256 units with ELU activation. ICM weights its forward and inverse losses by 0.2 and 0.8, respectively, and multiplies their weighted sum by 10. Its reward is half the sum of squared forward-prediction errors, multiplied by 0.01 and capped at 1.

E3B learns features through inverse dynamics and applies L2 normalization. It updates the feature precision matrix within each real episode and uses it to compute the elliptical exploration bonus, with a ridge regularization coefficient of 1. This matrix is neither an additional policy input nor reconstructed in imagined trajectories. Imagined rewards are therefore approximated by the reward head from the recurrent latent state rather than obtained through explicit matrix updates.

SMiRL converts each 64×64 image to grayscale and averages non-overlapping blocks to produce 8×8 features in $[0, 1]$. It estimates their density with a factorized Gaussian model. Each episode initializes the density from its first observation, regularizing the variance with unit variance and a pseudocount of 1. Subsequent observations are scored before the density is updated. The standard deviation is floored at 0.001, and the log density is clipped to $[-300, 300]$ and multiplied by 0.00125 to obtain the reward. The density's means, standard deviations, and count form 129 statistics provided to the encoder and used as decoder reconstruction targets. Their changes during imagination are predicted by the latent dynamics rather than explicit density updates at each step.

P2E predicts the next stochastic latent state from the current latent state and action using ten prediction heads, each with four layers of 400 units and ELU activation. Disagreement is the population standard deviation across the ten heads in each feature dimension, averaged over feature dimensions, with a reward coefficient of 1. The heads are trained on real transitions, and their auxiliary prediction loss does not update the world model.

The additional modules of RND, ICM, E3B, and P2E use separate Adam optimizers. The learning rate is 10^{-4} for the first three methods and 3×10^{-4} for P2E, with $\epsilon = 10^{-5}$. Global gradient norms are clipped at 40, 100, and 100 for ICM, E3B, and P2E, respectively. RND uses no gradient clipping.

Table C1: Training settings shared by all methods in Crafter.

Item	Setting
Gradient estimator for observation posterior	Gumbel straight-through, applied only to the observation posterior
Main optimizer learning rate	8×10^{-5}
Batch size / sequence length	16 / 64
Discount factor γ	$1 - 1/333 \approx 0.997$
Imagination horizon	15
λ-return parameter	0.95
Replay capacity	500,000 transitions
Training ratio	32
Parallel environments	8
Maximum training episode length	1,000,000 environment steps
Maximum final-evaluation episode length	10,000 environment steps

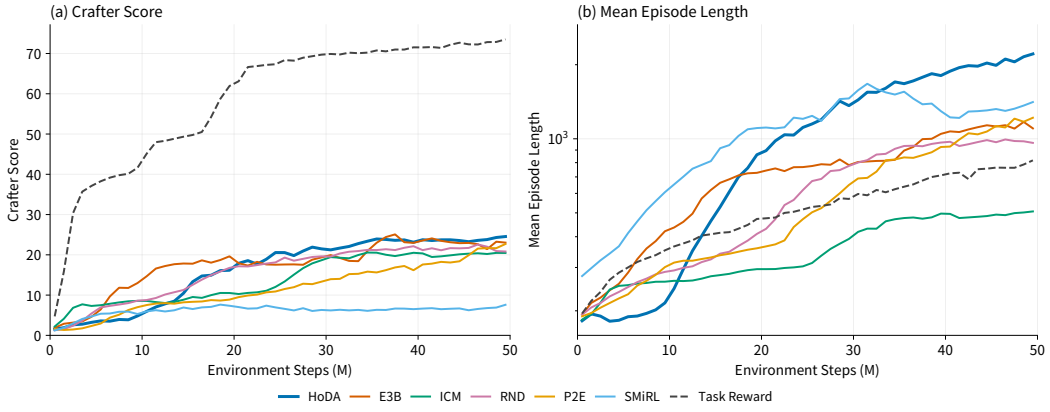

Figure D1: Crafter score and episode length during training. (a) Crafter score from periodic evaluation. (b) Mean length of episodes completed over the same period, on a logarithmic scale. The horizontal axis shows collected environment steps.

C.2 Shared Training Settings

All methods compared in Crafter use the 25-million-parameter backbone size of DreamerV3 (Hafner et al., 2025) and share the settings in Tab. C1.

The training ratio is the number of training-data steps per environment step collected. The DreamerV3 Crafter preset uses 512, whereas all methods in our Crafter experiments use 32.

In Crafter, final evaluation resets the environment and policy memory at the start of each episode. Episodes end at death or the 10,000-step limit, and all are included in the statistics.

D Crafter Results and Analysis Methods

D.1 Training Curves

The curves in Fig. D1 summarize results over intervals of one million environment steps. Within each training run, the left panel averages periodic Crafter scores in each interval, and the right panel averages the lengths of episodes ending in that interval. Complete runs of each method are then weighted equally.

D.2 Physiological State Statistics and Time Courses

Fig. 2(a) first averages each normalized physiological variable over time within each episode, then across episodes within each training run, and finally equally across runs. Fig. 2(b) computes D at each step using Eq. (2) and takes its episode time average. Points include evaluation episodes from all complete runs. Whiskers extend to the most extreme observation within 1.5 times the interquartile range. Fig. 2 uses states after real actions, includes the terminal step, excludes reset states, and does not fill values after death.

Fig. D2 shows a subset of the episode-level water, food, and energy trajectories summarized in Fig. 3 over the first 1,000 actions. For clarity, the heatmap displays three training seeds per method, including all 256 evaluation worlds for each selected seed. Runs are ordered first, and worlds retain the same fixed order within each selected run. Episodes are not selected or sorted by lifetime or behavioral outcome, and trajectories are not aligned to replenishment or sleep events.

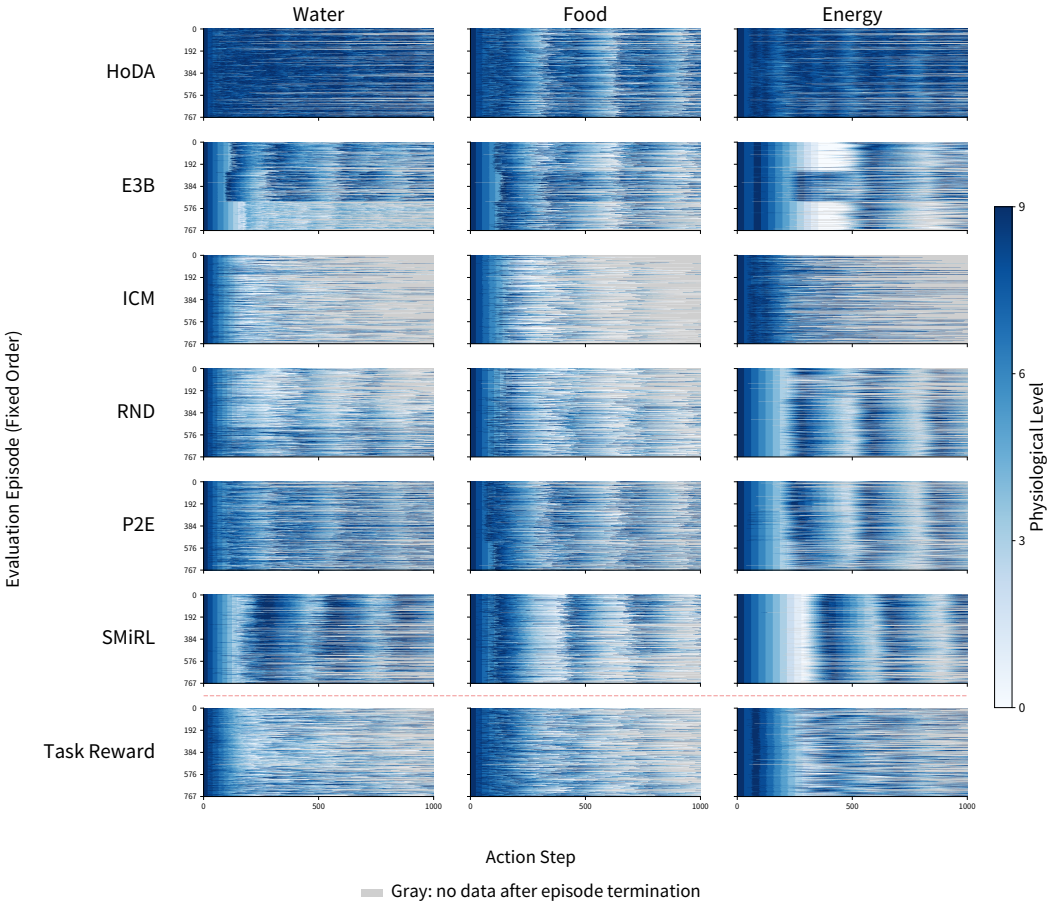

Figure D2: Episode-level physiological state. For visualization, we show a subset of three training seeds per method, including all 256 evaluation worlds for each selected seed. Each row is one episode, with the same row order across a method's three panels. Light-to-dark colors show values from 0 to 9 on a shared scale. Gray denotes the period after an episode has ended, not a physiological value of zero.

In Figs. 3 and D2, time 0 is the initial state before the first action. Later values are recorded after each real action, including the terminal step. At each step, Fig. 3 equally averages episodes with a recorded value and shows their 25th–75th percentiles. The band is not a confidence interval. The rightmost column gives the fraction of episodes with a recorded value. Values after episode end are neither filled with zero nor carried forward, and episodes are not preselected for surviving to the end of the window. Early depletion follows nearly identical discrete steps across training runs and

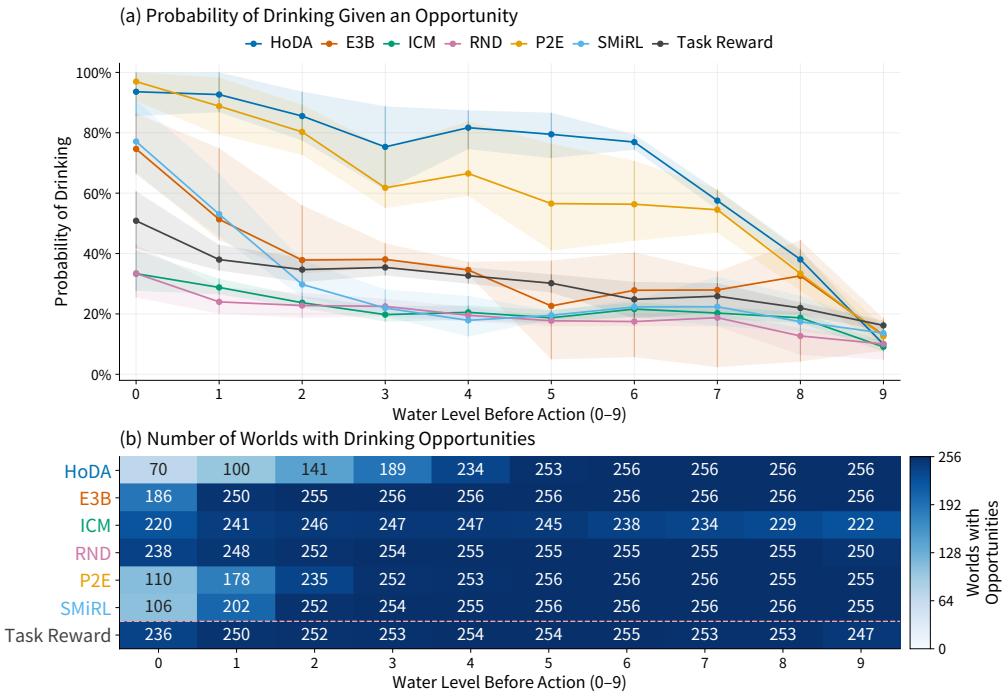

Figure D3: Drinking choices and current water level across methods. (a) Percentage of drinking opportunities in which the agent chooses to drink at each water level, with shading for pointwise 95% intervals. (b) Number of worlds supporting each estimate. Each world counts at most once per water level.

worlds. The resulting low between-episode variance makes the mean stepwise as well. Percentile bands are not smoothed, and their boundaries change in steps as quartiles cross integer levels.

Shared rhythms and environmental constraints. The episode-level heatmaps show that the repeated fluctuations in the aggregate curves arise from depletion and recovery across many episodes rather than a few unusually long ones. Crafter's day–night cycle repeats every 300 steps, and food availability and enemy threats vary with lighting. The physiological update rules also deplete food and energy with fixed timers and restore energy through sleep. Because every method faces these rules, their policies may learn similar replenishment and rest rhythms.

D.3 Drinking Choices and Current Water Level

To distinguish whether drinking occurred from whether the policy chose to drink at a given water level, we select evaluation states in which the agent is alive, awake, facing water, and the target cell is not occupied by an object before the action. In these states, selecting the environment's interact action allows the agent to drink. We group states by the pre-action water level from 0 to 9 and divide the number of times the agent selects the interact action by the number of available drinking opportunities in each group. Steps during automatic sleep and at the transition to wakefulness are excluded from awake opportunities. Intervals for the choice percentages use percentiles from stratified resampling over independent runs and shared worlds. A world with no drinking opportunity at a given water level contributes no observation, rather than a zero percentage. Resamples with no available opportunity are excluded from the interval estimate.

Fig. D3(a) shows that agents trained with every method choose to drink more often when water is depleted than when it is full, with the choice rate generally decreasing as water level rises. HoDA and P2E still choose to drink relatively often at intermediate water levels, whereas RND and ICM have lower, less variable choice rates in this range. Fig. D3(b) shows how many worlds contribute to each estimate. For example, HoDA's drinking opportunities at low water levels come from fewer worlds than those at high levels. Shared replenishment behavior across intrinsic drives thus extends beyond

Table D1: Percentage of episodes with shelter construction followed by sleep, a return to a shelter area followed by sleep, or path construction and use in water. Results use complete episodes unless marked as the first 1,000 steps.

Method	Shelter construction	Return and sleep (full episode)	Return and sleep (first 1,000 steps)	Path construction and use
HoDA	93.0 $\pm$ 4.3	28.4 $\pm$ 15.8	7.0 $\pm$ 3.7	92.1 $\pm$ 2.7
E3B	74.5 $\pm$ 9.7	0.9 $\pm$ 0.2	0.3 $\pm$ 0.2	86.1 $\pm$ 11.0
ICM	45.6 $\pm$ 4.6	1.0 $\pm$ 0.6	0.9 $\pm$ 0.8	37.6 $\pm$ 1.8
RND	78.5 $\pm$ 9.8	4.4 $\pm$ 2.0	1.4 $\pm$ 0.6	79.6 $\pm$ 9.7
P2E	73.7 $\pm$ 2.3	5.6 $\pm$ 3.9	1.4 $\pm$ 1.8	70.7 $\pm$ 5.6
SMiRL	34.8 $\pm$ 11.4	1.8 $\pm$ 1.8	0.9 $\pm$ 0.6	0.5 $\pm$ 0.9
Task Reward	49.1 $\pm$ 1.4	1.2 $\pm$ 0.8	0.7 $\pm$ 0.5	51.6 $\pm$ 3.6

simply drinking to drinking choices that are related to current bodily state. This relationship does not require rewards that directly measure water deficit. Agents with non-homeostatic intrinsic drives may also learn state-dependent replenishment choices through the effects of drinking on subsequent interaction.

D.4 Shelters and Paths through Water

Shelter measurement. We measure the percentage of episodes in which agents enclose an area and sleep inside it. We reconstruct the map action by action, tracking successfully placed stones, crafting tables, furnaces, and plants that remain present. At sleep onset, we identify the agent's walkable region using four-neighbor connectivity over the full map. A case qualifies if removing agent-built barriers in a copy of the map and restoring the pre-placement walkability reconnects this region to an external region with a larger area that the agent previously reached. Water and natural obstacles may form part of the boundary alongside agent-built barriers. Lava remains walkable, and moving animals or enemies are not treated as fixed barriers. Merely placing stone on water, sleeping in a naturally enclosed area, or resting in a partly enclosed area with an open route outside does not qualify. Each episode counts at most once, regardless of repeated sleep. Episodes with no sleep or an early ending remain in the denominator.

Shelter-return measurement. We measure the percentage of episodes in which agents leave an area they previously enclosed and slept in, then return to sleep there again. Both sleep events must meet the shelter criterion above, and both sleeping positions must lie inside both enclosures. Between the two sleep events, the agent must enter the larger, previously reached external walkable region separated by the first enclosure. That position must also lie outside the second enclosure. The agent may change its sleeping position within the area or close the entrance again. Movement only within the enclosure or repeated sleep without leaving does not count. Each episode counts at most once, and episodes without a return remain in the denominator.

Construction and use of paths through water. We count the complete sequence of successfully placing stone on a water cell, mining the stone to turn the cell into walkable ground, and then entering that cell. All three steps must occur in order at the same location. Placing stone alone, mining natural stone, or creating a path without traversing it does not count. Each episode counts at most once, and episodes without the complete sequence remain in the denominator. Tab. D1 summarizes the percentage of episodes with each of the three behaviors.

All methods sometimes return to a previously used shelter area to sleep again. This behavior is more frequent under HoDA, also occurs under P2E and RND, and is less common under the remaining methods. To examine the effect of interaction duration, Tab. D1 also reports the percentage of episodes with a return followed by sleep within the first 1,000 steps. A return counts only if the agent returns and sleeps again by the cutoff, and episodes that end earlier remain in the denominator. For most methods, this percentage is lower than over complete episodes, indicating that interaction duration affects the results. The timing of first shelter construction and subsequent choices about leaving and where to rest may also affect how often such returns occur. Their specific roles require further study.

Constructing and using paths through water is common under HoDA, E3B, RND, and P2E, also occurs under ICM, and is rare under SMiRL. Together with the shelter results, this suggests that different intrinsic drives can support terrain modification for movement and rest without necessarily producing the same construction patterns. Paths through water can also provide access to shelters.

Construction example. Fig. 4 shows agents from each method sleeping in shelters that meet the criterion above. SMiRL uses a crafting table as part of its enclosure, while the other examples use agent-placed stones together with water, trees, or rock walls. Uniform lighting is applied only for display. The scene and sleep state are unchanged, as are the observations actually received by the agent.

Fig. D4 shows the full construction sequence for HoDA's water-based example from Fig. 4. The agent first places a stone on a water cell and then mines the placed stone, converting that cell from water into passable ground. Repeating the operation farther into the water creates a continuous path. After entering the path, the agent returns to the inner side of the entrance, places a stone to close it, and sleeps inside the enclosure.

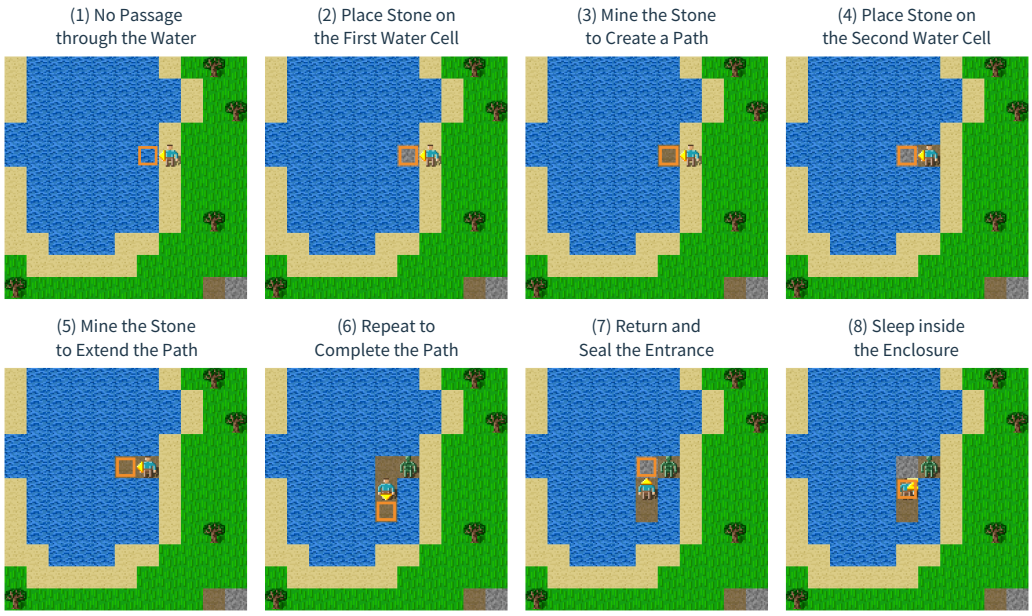

Figure D4: Construction of a water-based shelter. Orange boxes mark the key cell changed or confirmed at each step.

D.5 Rewards and Action Choice during Shelter Construction

Agents trained with different intrinsic drives all construct shelters. We examine how rewards received when construction actions are executed relate to action choice, with Task Reward as a reference. Intrinsic-drive methods use their own intrinsic rewards. Task Reward uses rewards for first-time achievements within each episode and health changes. We analyze successful placements of stones, crafting tables, and other objects that later form part of the boundary of an enclosed sleeping area. These actions are not limited to the final placement that closes the entrance. Each reward is expressed as its percentile rank among rewards during awake steps in the same episode.

For the matched comparison in Fig. 5, we use successful placements in the same world that do not contribute to shelter construction as controls. Matched actions must have the same action type, occur within 300 steps of each other, and differ by at most 0.2 in light level and 2 in each physiological variable. Within each episode, we average differences in reward rank and action probability across valid pairs, subtracting ordinary placement from shelter placement. We then average equally over contributing episodes within each training run and equally across runs, and estimate 95% confidence intervals by resampling.

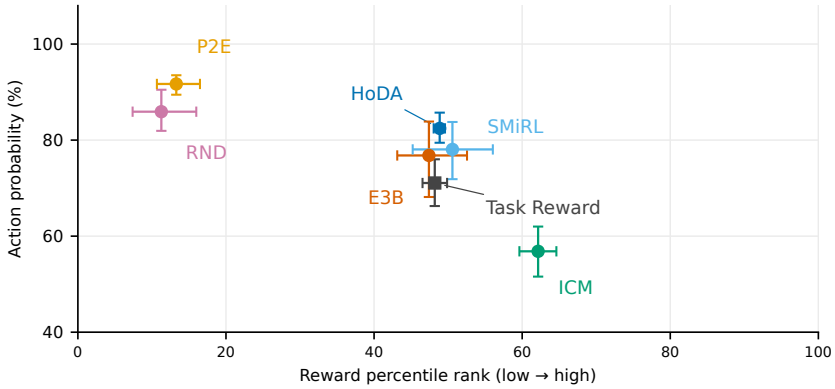

Figure D5: Immediate rewards and action probabilities during shelter construction. Points farther left indicate lower within-episode reward percentile ranks for shelter-placement actions. Points higher up indicate higher probabilities assigned to these actions. Horizontal and vertical error bars show the respective 95% confidence intervals. Task Reward has many tied zero rewards, whose percentiles are computed using midpoint ranks.

Unlike the differences shown in Fig. 5, Fig. D5 shows the reward ranks and action probabilities of shelter-placement actions themselves. For both metrics, we first average over placement actions within each episode, then average equally across contributing episodes within each training run and equally across runs, excluding episodes without qualifying actions.

Fig. D5 shows that RND and P2E assign high probabilities to shelter-placement actions despite their low reward percentile ranks. These actions have higher reward percentile ranks under ICM, while HoDA, E3B, and SMiRL lie in the middle range. Similar construction behavior therefore does not correspond to a common level of immediate reward.

Task Reward also assigns higher probabilities to shelter-placement actions without clearly higher task rewards at those steps than for ordinary placements. Crafter does not directly reward shelter construction, although individual placements may unlock achievements. Task Reward also includes explicit health feedback. By reducing later injuries and supporting health recovery, shelters may directly improve returns, while survival may allow more achievements to be completed. Thus, zero reward at a construction step does not rule out a role for health rewards in this behavior.

The ordinary-placement comparison helps interpret ICM's higher reward percentile ranks. Changes caused by object placement may be difficult to predict, yielding higher prediction-error rewards (Pathak et al., 2017), but ICM does not assign clearly higher rewards to placements that contribute to shelters. Its higher prediction error therefore should not be interpreted as a reward specifically for shelter protection.

E3B measures novelty using features of states visited within the current episode (Henaff et al., 2022). States encountered during training may still yield exploration rewards in a new episode, so E3B need not share RND and P2E's low reward levels. SMiRL instead rewards observations with high likelihood under its density model (Berseth et al., 2021). Object placement changes the current image, whereas a shelter may reduce disturbances during later sleep. The reward during construction can therefore differ from the later benefit of maintaining predictable observations. Both methods' positions in the figure also depend on visitation history and current state, not only on the reward definition.

For HoDA, we further examine whether physiological rewards arise mainly during shelter construction or during recovery while sleeping inside it. Fig. D6 compares the mean per-step reward contributions from food, water, energy, and health changes in these two phases, distinguishing construction costs from recovery benefits during use.

Fig. D6 shows that, during object placement for shelter construction, negative contributions from food, water, and energy depletion outweigh the positive contribution from health recovery, producing a negative mean net reward. During sustained sleep inside the shelter, energy and health recovery

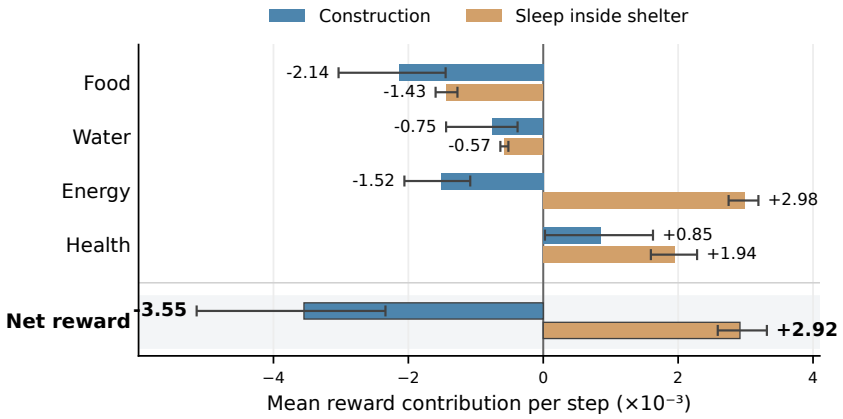

Figure D6: HoDA's mean per-step rewards for shelter-placement actions and sustained sleep inside shelters. Each row compares the two behaviors, with negative contributions to the left and positive contributions to the right. The final row sums the four physiological contributions to give the net reward for each behavior. Error bars show 95% confidence intervals.

Table D2: Spatial visitation and episode length.

Method	**Mean unique cells visited**	**Mean episode length (steps)**
HoDA	540.20 ± 162.13	2232.23 ± 642.63
E3B	331.02 ± 43.01	1188.48 ± 234.96
ICM	82.51 ± 3.17	527.54 ± 21.80
RND	161.66 ± 49.40	1017.34 ± 266.57
P2E	321.68 ± 17.64	1131.51 ± 50.16
SMiRL	208.11 ± 63.07	1090.65 ± 241.48
Task Reward	221.53 ± 33.82	844.92 ± 126.13

instead offset food and water depletion, making the mean net reward positive. HoDA thus incurs costs during construction and receives recovery rewards during use.

Overall, construction actions receive different immediate rewards across methods, yet become part of the learned policies. Similar shelter behavior may serve bodily recovery, continued exploration, or predictable observations without requiring similar rewards at the time of construction. These explanations are consistent with the results, but the analysis concerns learned behavior along each method's own trajectories. It does not establish how these benefits led to skill formation during training.

D.6 Spatial Visitation and Survival

Unique cells visited are counted from the player's position before each action and include the initial position. A landing cell reached only after the terminal action is not counted separately. For Fig. 6, we record whether each coordinate is visited at least once in each episode, average over evaluation episodes within each independent run, and then weight runs equally. Repeated visits do not increase an episode's contribution, so a cell value is the fraction of episodes that visit it, not a dwell-time distribution. Summing cell probabilities gives the mean number of unique cells visited per episode. All seven panels use the same square-root color scale, with the colorbar labeled in the original probability scale. This display makes low-probability regions more visible, but equal color differences do not represent equal probability differences.

Tab. D2 reports the mean number of unique cells visited and mean episode length over complete episodes. HoDA is highest on both measures. E3B and P2E also visit many cells. SMiRL covers fewer cells, while ICM visits the fewest. These results complement the physiological maintenance results. Intrinsic drives that encourage exploration also support replenishment and rest, while HoDA

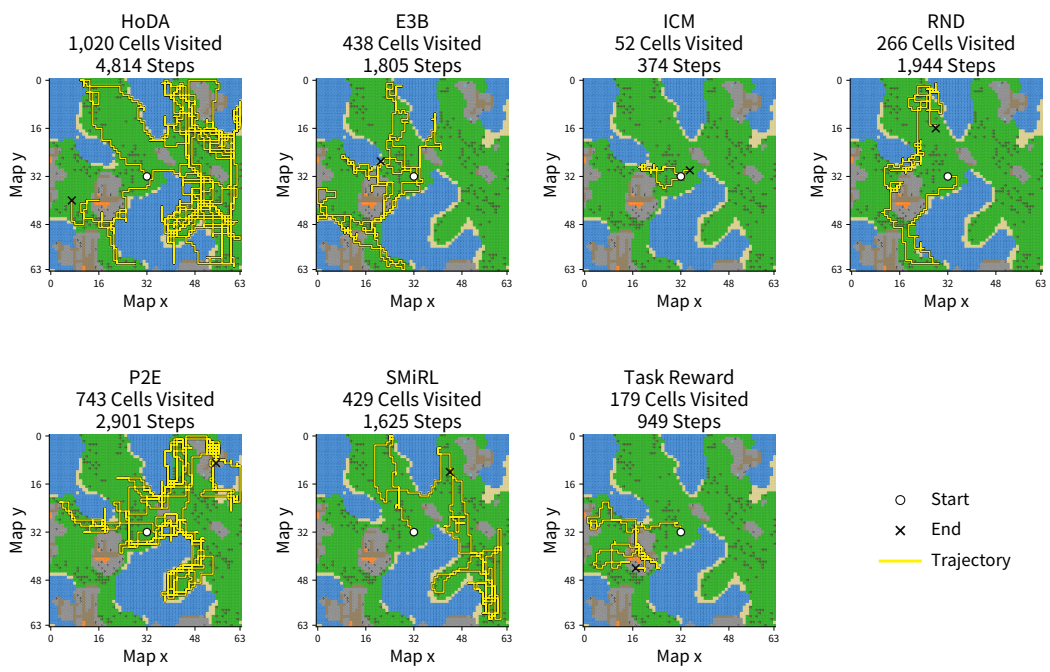

Figure D7: Single-episode trajectories of all methods in the same world. Cell counts include each visited cell only once.

and SMiRL show spatial exploration without novelty or map-coverage rewards. Different intrinsic drives thus support similar skills.

Fig. D7 shows trajectories in one shared world. Several intrinsically driven agents reach multiple regions beyond the starting area. HoDA has the broadest coverage in this example, extending from the center to several separated major land regions around the map. HoDA does not directly reward visits to new locations. One possible explanation is that bodily resources are continuously depleted while supplies are scattered, making long-range movement useful for finding resources, handling threats, and staying alive. Exploration may thus become a means of meeting bodily needs.

Exploration and survival at matched action budgets. Tab. D2 uses complete episodes, so agents that interact longer also have more opportunities to explore. To separate exploration speed from interaction duration, Fig. D8 compares the number of unique cells visited at matched action budgets and simultaneously shows the fraction of episodes still active. After an episode ends, its cumulative number of unique cells visited remains fixed and continues to contribute to the mean.

Fig. D8 shows that E3B and HoDA visit similar numbers of unique cells on average within the first 1,000 steps, but more of HoDA's episodes remain active, allowing coverage to continue expanding afterward. P2E's early count of unique cells visited is also close to HoDA's, but its episodes end sooner. HoDA's broad exploration thus accompanies longer survival. This suggests that staying alive may indirectly support broader coverage by allowing more opportunities to act.

D.7 Sensitivity to HoDA's Setpoint

The main experiments use a setpoint of 1.0 for all four physiological dimensions. To test how this choice affects physiological rhythms and training performance, we additionally compare setpoints 1.0, 0.8, and 0.5 while holding all other training settings fixed. Physiological time courses use the shared world list and cover the first 1,000 actions. At each step, only episodes with a recorded value contribute. Values after termination are not filled, and curves are not smoothed.

Water, food, and energy under setpoint 0.8 remain broadly similar to the main setpoint of 1.0. Setpoint 0.5 produces the clearest energy depletion–recovery cycle, but episodes also end sooner. Changing only the target physiological level within the same homeostatic reward produces energy

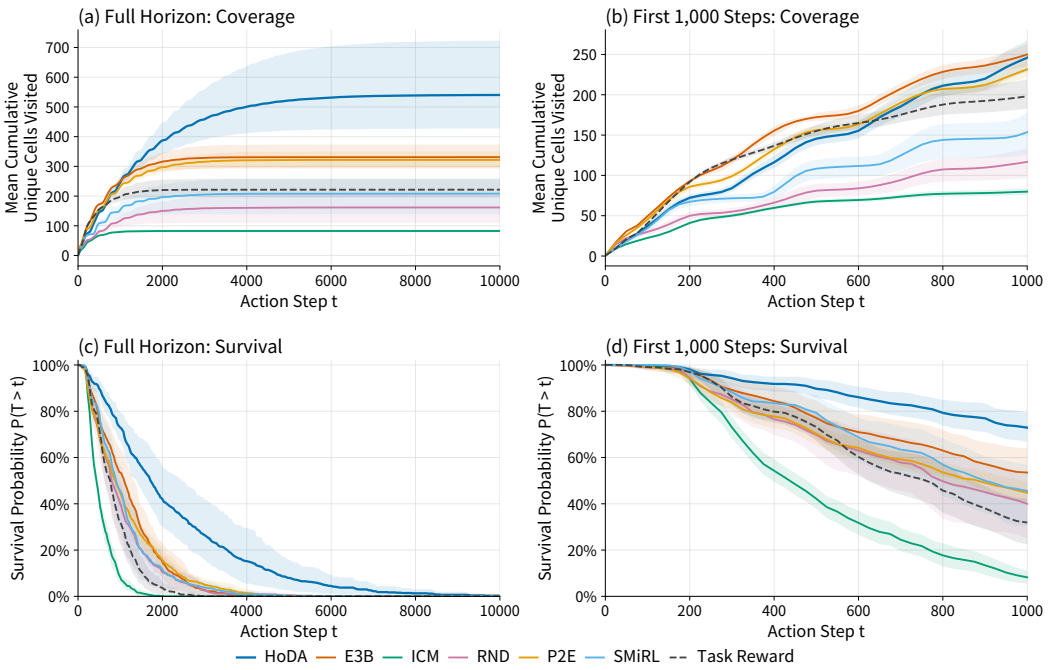

Figure D8: Cumulative unique cells visited and percentage of episodes still active at matched action budgets. The top row shows the mean number of unique cells visited, and the bottom row shows the percentage of episodes still active. The left column shows the full time course, and the right column enlarges the first 1,000 steps. Shading denotes 95% intervals from resampling independent runs and shared worlds.

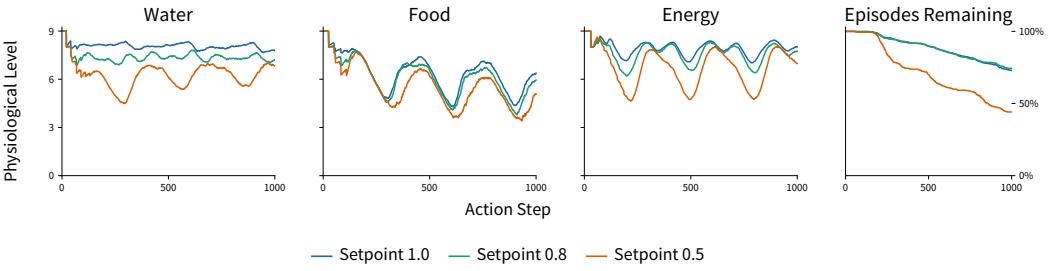

Figure D9: HoDA's physiological time courses under three setpoints. The first three columns show mean water, food, and energy levels over episodes with recorded values, and the right column shows the percentage of episodes contributing.

cycles closer to those of other intrinsic methods. Food cycles remain broadly aligned across all three setpoints, which may reflect the same environmental cycle.

Tab. D3 shows that, within the 50M training budget, setpoint 0.8 yields a higher Crafter score than 1.0 and a similar mean episode length, while setpoint 0.5 yields lower values for both metrics.

Table D3: Crafter score and mean episode length of the final models under three setpoints.

Setpoint	**Crafter score**	**Mean episode length (steps)**
1.0	27.77 $\pm$ 1.38	2232.23 $\pm$ 642.63
0.8	29.80 $\pm$ 0.24	2208.08 $\pm$ 20.56
0.5	22.54 $\pm$ 1.61	1111.70 $\pm$ 52.19

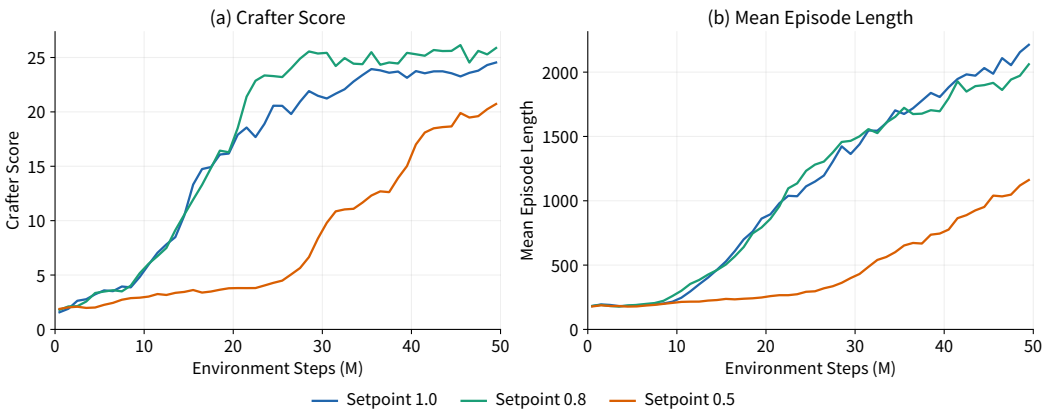

Figure D10: Training curves of HoDA under three setpoints. Curves aggregate Crafter score and the length of completed episodes in one-million-step windows during training.

Table D4: Final evaluation of alternative homeostatic reward forms. Achievement columns report the percentage of episodes that complete the corresponding achievement at least once.

Reward form	**Crafter score**	**Mean achievements**	**Mean episode length (steps)**	**Collect iron (%)**	**Craft stone pickaxe (%)**	**Craft stone sword (%)**
HoDA, $D_{t-1} - D_t$	27.77	13.60	2232.23	6.12	24.87	92.84
Negative-drive reward, $-D_t$	2.16	4.04	147.10	0.00	0.00	0.00

Fig. D10 shows that both metrics at setpoint 0.5 continue to improve late in training. Together with Fig. D9, these results show that the setpoint affects both physiological rhythms and learning progress. At setpoint 0.8, iron pickaxes are also crafted in 0.20% of evaluation episodes, but iron-tool crafting remains rare rather than reliable.

D.8 Homeostatic Drive-Reduction Reward versus Negative-Drive Reward

To test HoDA's sensitivity to its reward formulation, we compare HoDA's drive-reduction reward with an alternative that directly uses negative homeostatic drive, $-D_t$. The former evaluates improvement in physiological state, whereas the latter penalizes the current homeostatic drive. Both agents complete 50 million training steps and differ only in reward form. Tab. D4 uses the final-evaluation protocol of Section 4.2 to compare the two reward forms under the same conditions.

Under the same training and final-evaluation setup, the agent using homeostatic drive-reduction rewards survives longer and learns more skills. Reward form therefore affects learning even with the same physiological objective. The two rewards produce different return distributions, so normalization may affect learning differently. Because negative-drive rewards are nonpositive at each step, they may encourage early termination if ending the episode avoids future negative returns. The relation between the two returns also includes different terminal boundary terms (Appendix B.3).

E Behavioral Analysis in Atari

E.1 Experimental Setup

To examine whether shared behaviors also emerge beyond Crafter, we compare RND, P2E, ICM, E3B, SMiRL, and Empowerment in Seaquest and Ms. Pac-Man, alongside a uniform random-action baseline (Random). HoDA is evaluated only in Seaquest, using normalized oxygen level as its physiological variable with a setpoint of 0.5. Its reward follows the reduction in squared physiological deficit defined in Section 3.2. Rewards are set to zero at game start, upon life loss,

during explosion animations, and at episode termination. The first underwater observation after each such boundary initializes the drive, and reward calculation resumes on the next step, excluding reset-related oxygen increases. All trained methods use the DreamerV3 backbone and learn only from intrinsic rewards. Native game scores are used only for evaluation and event detection.

The Atari experiments use the 12-million-parameter backbone size, a main optimizer learning rate of 4×10^{-5}, 16 parallel environments, and a replay capacity of 200,000 transitions. The batch size is 16, sequence length is 64, training ratio is 32, and imagination horizon is 15. Training and evaluation steps are counted as agent decisions, each corresponding to four game frames. We evaluate each learned method after 50M decision steps (200M game frames) of training.

Unlike the Crafter configuration, Atari SMiRL uses a reward scale of 0.015625 and estimates variance directly from within-episode observations, without unit-variance pseudocount regularization. Empowerment uses a variational mutual information approximation under the current policy. Along imagined trajectories, an auxiliary network predicts the current action from the current latent state and the stochastic latent variable five steps later. The intrinsic reward is the predicted log probability of that action minus its log probability under the policy. Policy learning uses imagined rewards without a value loss on real replay data.

We evaluate all methods using the same environment initialization seeds and sample actions from the policy distribution. Episodes are evaluated for up to 5,000 steps, with the percentage of episodes exhibiting each behavior reported for both the first 1,000 steps and the full evaluation period. Inputs are 96×96 grayscale images. Each action is repeated for four frames, sticky actions are enabled (Machado et al., 2018), and up to 30 random no-op actions are applied at the start. Losing a single life does not end an episode.

E.2 Surfacing for Oxygen and Diving Again in Seaquest

Underwater activity in Seaquest consumes oxygen, which can be replenished by surfacing. We measure the percentage of episodes in which the agent spends time underwater, surfaces to replenish oxygen, and dives again within the same life. We also measure how often this cycle occurs at least twice in an episode. A valid cycle requires at least 16 underwater steps before replenishment, an oxygen increase of at least four units at the surface, and a return underwater before losing the current life. Oxygen changes due to life resets and explosion animations are excluded. The agent need not be close to running out of oxygen before replenishment.

Table E1: Percentage of episodes with oxygen replenishment followed by diving again in Seaquest. The last two columns report completion at least once and at least twice within the 5,000-step evaluation limit.

Method	Completed within 1,000 steps	Completed at least once	Completed at least twice
HoDA	19.9	19.9	1.2
RND	99.7	99.7	95.6
P2E	98.8	99.6	95.1
ICM	11.7	45.7	23.0
E3B	99.2	99.2	94.1
SMiRL	0.0	0.0	0.0
Empowerment	98.8	99.6	96.9
Random	13.3	13.7	1.2

Tab. E1 shows that, over the full evaluation period, all trained methods except SMiRL complete the oxygen replenishment and diving cycle in a larger fraction of episodes than random actions. RND, P2E, ICM, E3B, and Empowerment also repeat this cycle more often than random actions. Several intrinsic drives thus support similar behavior that restores a resource and allows activity to continue. By enabling further underwater activity, oxygen replenishment may preserve opportunities to obtain future intrinsic rewards under different drives.

E.3 Power-Pellet Collection and Ghost Eating in Ms. Pac-Man

Power pellets in Ms. Pac-Man temporarily make ghosts edible. We measure the percentage of episodes with power-pellet collection, subsequent ghost eating, and eating at least two ghosts during the effect of a single power pellet. Power-pellet events are identified from both the disappearance of the corresponding object and score changes. Ghost-eating events require contact between the player and a ghost during the active period, together with the corresponding score change. Life changes and fruit scores are excluded.

Table E2: Percentage of episodes exhibiting each behavior in Ms. Pac-Man. Except for the column explicitly limited to the first 1,000 steps, all columns cover up to 5,000 evaluation steps. The last column requires eating at least two ghosts during the effect of a single power pellet.

Method	Collected a power pellet	Ate a ghost within 1,000 steps	Ate a ghost	Ate multiple ghosts per power pellet
RND	97.3	71.5	90.6	57.3
P2E	99.3	81.2	86.2	35.9
ICM	98.4	60.2	62.5	20.7
E3B	99.6	61.3	86.3	35.9
SMiRL	97.7	77.7	78.9	5.1
Empowerment	0.0	0.0	0.0	0.0
Random	7.0	5.5	5.5	0.8

Tab. E2 shows that agents trained with RND, P2E, ICM, E3B, and SMiRL all collect power pellets and subsequently eat ghosts. Under each method, agents also eat multiple ghosts during a single power pellet's active period, and ghost eating occurs within the first 1,000 steps. All three behaviors occur less frequently under random actions. These five intrinsic drives thus support similar ghost-eating behavior, whereas agents trained with Empowerment neither collect power pellets nor eat ghosts.

Together, the two games show that different intrinsic drives can support similar functional behaviors, but these similarities depend on the environment and behavior. SMiRL collects power pellets and eats ghosts in Ms. Pac-Man but does not complete the oxygen replenishment and diving cycle in Seaquest. Empowerment shows the opposite pattern. These results support connections at the behavioral level, rather than identical policies or capabilities across all drives.