BUILDING WORLD MODELS FROM INTRINSIC MOTIVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In the absence of extrinsic rewards or task supervision, human behavior is guided by intrinsic motivation—neural signals that drive curiosity, creativity, play, and other essential facets of intelligence. Intuitively, intrinsic motivation holds a fundamental role in developing robust internal models of the world that enable generalization to new tasks from little to no task-directed experience. However, this hypothesis has not been carefully explored: existing approaches often focus on model-free reinforcement learning, compare only a small number of intrinsic motivations, use inconsistent reinforcement learning algorithms, and fail to isolate the benefits of representation learning from those of policy learning. This makes it impossible to compare different hypotheses for intrinsic motivation, and more importantly, ignores the interplay between intrinsic motivation and world-model building. To close these gaps, we integrate 24 state-of-the-art self-supervised exploration algorithms—spanning a breadth of methods across curiosity, entropy, and skill-discovery—with a common model-based reinforcement learning backbone, and evaluate how well these intrinsic motivations learn an effective world model. We test our agents on zero-shot, few-shot, and multi-task control tasks in locomotion, manipulation, and open-ended gaming environments using three competing world-model representation learning hypotheses: observation reconstruction, latent prediction, and redundancy reduction. Across all experiments, we find that curiosity-driven exploration with observation reconstruction has the strongest aggregate family profile across representations, but the best individual pairing can depend on the specific experiment. Reconstruction is particularly effective for visual zero-shot control, while the reconstruction-free models are achieve competitive transfer and few-shot adaptation. These findings motivate evaluating intrinsic drives by the reusable world models they build and designing intrinsic motivation and representation learning jointly. Our approach demonstrates that world models provide a powerful tool to measure the quality of intrinsically motivated experience, and identify intrinsic motivation as a fundamental ingredient for building world models just as animals do—autonomously.

1 INTRODUCTION

Humans and animals demonstrate a remarkable capacity to create and pursue intrinsic goals, adapt to an ever-changing environment, and learn new skills with a little practice. How do we achieve this feat of autonomy? One promising substrate is an internal model of the world, allowing us to reason and plan in imagination using relevant abstractions. This hypothesis perhaps originated with Kenneth Craik in *The Nature of Explanation* (1943):

> *"If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it."*

The kind of model Craik describes is a generative one: simulating the consequences of actions on future states, ultimately enabling planning and reasoning about the world in imagination. Such models

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

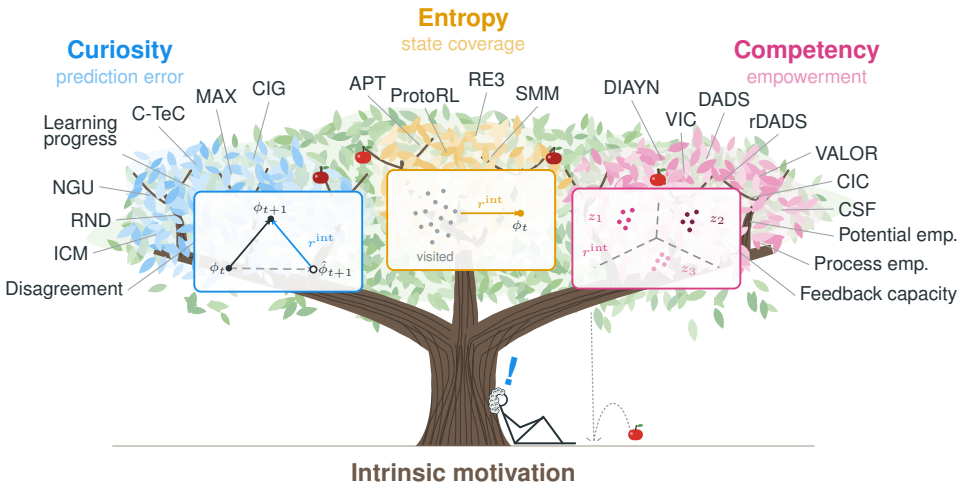

Figure 1: **Three families of intrinsic motivation.** Our comparison spans curiosity (blue), entropy (orange), and competency (pink). Leaves illustrate representative objectives and estimator families; insets illustrate prediction error, state novelty, and skill discriminability. These families overlap: NGU, for example, combines lifelong curiosity with episodic novelty. Here ϕ_t denotes a state representation, r^{int} intrinsic reward, and z_i a skill or action.

have been a longstanding topic of research in reinforcement learning (RL), culminating in the development of sophisticated algorithms like DreamerV3 and TD-MPC2 that learn complex behaviors in domains ranging from locomotion to Minecraft (Hafner et al., 2025; Hansen et al., 2024). The RL communitiy typically approaches model-free and model-base algorithms as a methodological choice, and model-free approaches are far more popular by virtue of their simplicity and scalability (**??**). However, recent theoretical work connects sufficiently general goal-directed behavior (including intrinsic goals!) with implicitly learning a predictive model, under formal assumptions about the goals and environments an agent can solve (Richens et al., 2025; Nayebi, 2026). From our perspective, understanding how world models are built up from experience is a fundamental question in the study of intelligence.

In this paper, we propose that intrinsic motivation serves a primary function in developing world models that generalize to new tasks from little to no task-directed experience. Therefore, a window into intelligent, intrinsically motivated behavior is the quality of the world model it builds for downstream control. State coverage and transfer return provide useful measurements, but neither alone specifies what an agent should learn about its world. Experience is valuable insofar as it supports a model that can predict the consequences of actions and guide future behavior. Studying the interplay between intrinsic motivation and world-model learning therefore offers a way to compare hypotheses about the purpose of exploration, as well as a practical route to better pretraining objectives.

This perspective builds on active world-model learning and Plan2Explore (P2E), which already demonstrate the value of curiosity for learning models before a task is known (Kim et al., 2020; Sekar et al., 2020). URLB compares several intrinsic-motivation families in a common model-free framework (Laskin et al., 2021), while LEXA learns visual goal-reaching behavior through a world model and a particular exploration strategy (Mendonca et al., 2021). Our contribution is to bring these two lines of inquiry together: a broad comparison of intrinsic objectives within one model-based framework, crossed with competing representation-learning objectives. To our knowledge, this is the first study at this breadth to compare 24 intrinsic objectives across three Dreamer representations while separating world-model transfer from policy transfer. Appendix A provides a fuller discussion of related work.

Challenges Intrinsic rewards depend on the representation in which novelty, prediction error, or skill discriminability is measured. Owing to DreamerV3's impressive ability to learn behavior in imagination across a wide breadth of domains under a single hyperparameter configuration, we use Dreamer's world model as a reasonable hypothesis for what counts as a "good" representation to

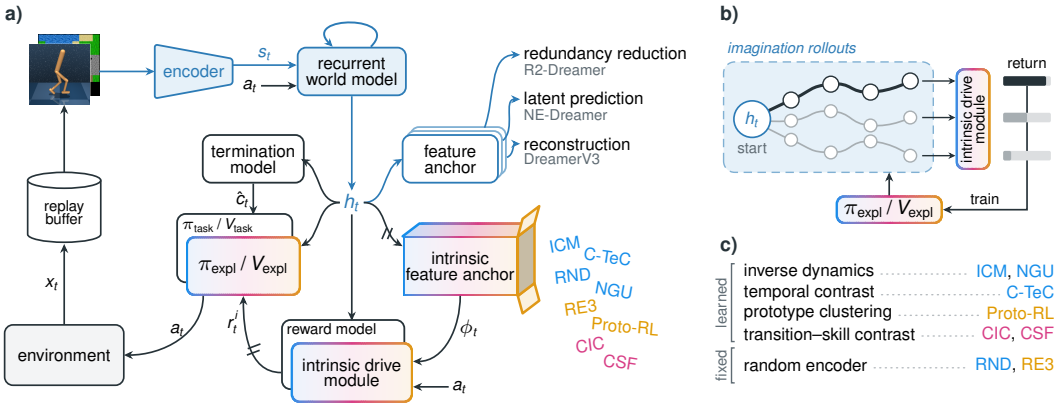

Figure 2: **Intrinsically motivated Dreamer agents.** (a) Exploration experience trains a shared recurrent world model using reconstruction (DreamerV3), next-embedding prediction (NE-Dreamer), or current-embedding redundancy reduction (R2-Dreamer). The intrinsic drive module (IDM) scores latent states or method-specific intrinsic features. Double slashes denote stopped gradients. (b) Intrinsic rewards on imagined trajectories train the exploration actor–critic. Separate task heads use predicted task rewards for evaluation. (c) Feature spaces used by representative IDMs.

unify different intrinsic motivation algorithms. In addition, the sheer diversity in algorithms we consider (24), especially with three different Dreamer backends, limits the number of environments we can test under reasonable compute. Thus, we chose our experiments to cover a few relevant high-level domains that still allow us to glean meaningful insight (physics-based control vs. a more abstract and cognitive game, continuous vs. discrete control, vision vs. proprioception). Finally, each algorithm has its own hyperparameters, and we do not attempt to tune them for each environment. Instead, we use the original hyperparameters from the source papers (taking into account author recommendations per environment when possible). This is a limitation of our study, but it also reflects our approach in using a breadth of algorithms within each family, where the many variants within a family capture some relevant differences in hyperparameterization.

Contributions To summarize, our contributions are as follows:

- We integrate 24 intrinsic objectives spanning curiosity, entropy, and competency into a common JAX Dreamer framework (Figure 1).
- We compare observation reconstruction with two reconstruction-free alternatives: next-embedding prediction and current-embedding redundancy reduction. We evaluate zero-shot control, multi-task transfer, and few-shot adaptation, distinguishing the experience selected by an intrinsic drive from the task heads trained to exploit it.
- Across these comparisons, curiosity provides the strongest overall family profile, but the benefits of pretraining depend on the representation and modality. These results make world-model learning an explicit mechanism for evaluating intrinsic motivation.

2 Methods

Existing studies of intrinsic motivation typically learn representations end-to-end by model-free policy optimization, such as SAC and PPO. This makes it difficult to understand how the efficacy of intrinsic motivation depends on the representation induced by the RL algorithm versus the intrinsic objective itself. In addition, it prevents us from studying the value of pretraining from intrinsic motivation on zero-shot control and transfer. Thus, we adopt a model-based approach using the basic recipe of Plan2Explore (Sekar et al., 2020) with DreamerV3's base world model as the common backbone. DreamerV3 has demonstrated remarkable generality across a wide range of tasks, and its recurrent state-space model (RSSM) provides a rich latent representation for intrinsic motivation. This makes it a particularly suitable algorithm as a common substrate to unify different intrinsic motivations in a feature space that has demonstrated success in offline planning. The flexibility

of Dreamer allows us to test two alternative feature anchors that implement distinct representation learning objectives (Figure 2). We provide a brief background on these algorithms as well as some design choices integrating them with intrinsic motivation.

2.1 Dreamer Backbone and Representation Learning

Dreamer learns an RSSM from replay and trains behavior through imagined trajectories. Givenan observation x_t, the encoder produces e_t, and the RSSM combines deterministic memory h_t with a stochastic state z_t to form a world-model state $s_t = (h_t, z_t)$. The posterior conditions on the current observation during training while the prior predicts latent states during imagination. All three backbones retain the same recurrent latent dynamics, continuation prediction, and the Dreamer actor–critic architecture, but differ in how observations anchor the latent state.

DreamerV3 *reconstructs the current observation* with a decoder, balancing the preservation of sensory information against KL regularization that encourages predictable latent representations (Hafner et al., 2025). NE-Dreamer predicts the *next encoder embedding* from latent-action history using a causal transformer (Bredis et al., 2026). R2-Dreamer aligns a linear projection of the *current latent state* with the current encoder embedding (Morihira et al., 2026). Both reconstruction-free variants use redundancy reduction to align paired features balancing against collapse (Zbontar et al., 2021). Thus, the comparison places sensory reconstruction alongside two forms of latent-space learning, including temporal prediction in the spirit of joint-embedding predictive architectures (JEPA; Assran et al., 2023). The specific auxiliary predictors and loss scales also differ between backeneds, so we emphasize the comparison as one between representation strategies rather than an isolated loss ablation.

We use separate exploration and task actor–critics, driven by intrinsic and predicted task reward, respectively. Task rewards can train a detached reward predictor and task actor–critic for evaluation, but their gradients do not update the world model or exploration policy. Accordingly, *zero-shot* and *few-shot* in our paper refers to task-directed data collection. During supervised training and online adaptation phase, task reward does train the shared representation.

2.2 Intrinsic Drive Module

We adapt each intrinsic motivation through an IDM that learns from replay and scores imagined transitions. All IDMs use the deterministic RSSM state component h_t as input,

$$\phi_t = f_\theta(\mathrm{sg}(h_t)), \qquad r_t^{\mathrm{int}} = g_\omega(\phi_t, a_t, \phi_{t+1}; \mathcal{M}_t), \tag{1}$$

where f_θ is an optional intrinsic feature mapping, sg is a stopgrad operator, r_t^{int} is the intrinsic reward with parameters ω, and $\mathcal{M}_t$ contains algorithm-specific state, such as episodic memory or a sampled skill. Predictors, density estimators, and discriminators that consintute a given intrinsic motivation all train on posterior trajectories (i.e. latent representations of real experience from replay) and their computed rewards train the exploration actor–critic on the imagined trajectories (from the prior). Skill-discovery methods additionally condition the exploration policy on the skill sampled from a skill prior. We block IDM training gradients at the RSSM and stop gradients through intrinsic rewards during actor updates. Each drive therefore shapes the world model through the experience it collects, without directly changing its representation loss.

Several methods require a particular feature space as part of their original formulation. We therefore allow an additional representation above the Dreamer state; for example, ICM learns inverse-dynamics features, whereas RND uses a fixed random projector (Fig. 2). Methods without this requirement use RSSM features directly. This retains important representational assumptions of the original algorithms while sharing a common world model. To the best of our ability, we implement the specific architectures, reward normalization, and relevant hyperparameters from their original papers and suggestions rather than imposing one setting on every intrinsic objective. In addition, our comparison includes some idiosyncratic variants of existing algorithms, such as RDADS and Vanilla Curiosity. Appendix C documents all 24 objectives and any deviations from their original implementations. Finally, we point out that our study includes several underexplored (e.g. NGU, SMM, Empowerment) or relatively new (e.g. CIG, C-TeC) algorithms, providing the first demonstration of their successes and failures in these domains and experiments.

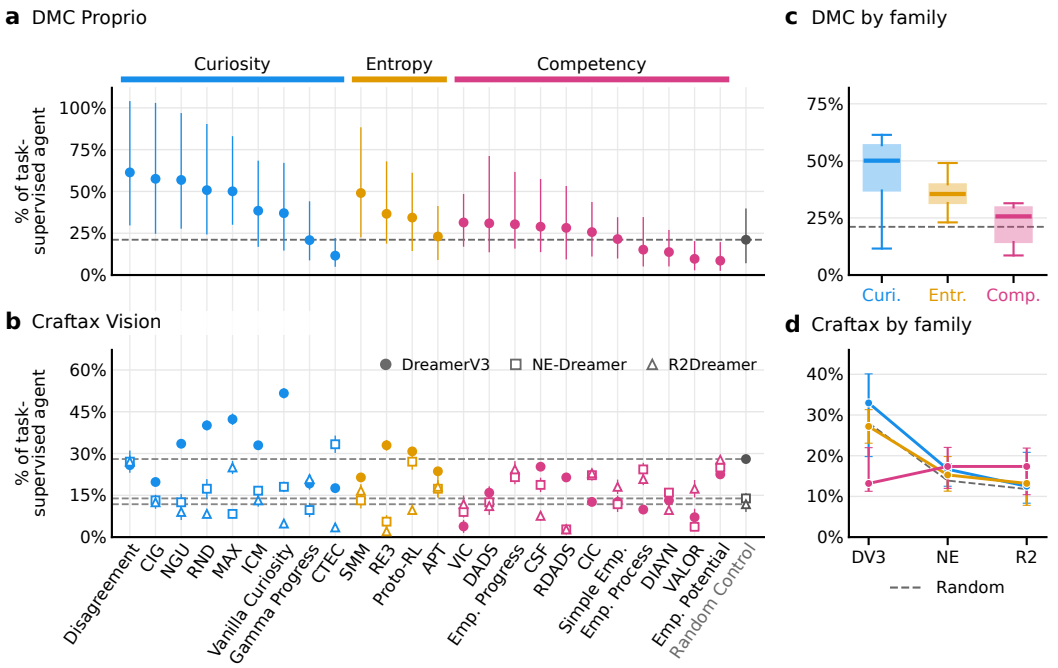

Figure 3: **Single-task zero-shot generalization.** Returns are normalized by the task-supervised reference. (a) DreamerV3 interquartile-mean (IQM) return across 25 proprioceptive DMC tasks after 1M exploration steps; error bars bootstrap tasks. (b) Craftax vision after 5M steps across three backbones. (c) DMC family distributions. (d) Craftax family medians and interquartile ranges across methods. Dashed lines denote Random Control.

2.3 Environment Choices and Experiment Design

Continuous control We evaluate all 25 tasks from the DeepMind Control (DMC) suite from proprioceptive observations using DreamerV3 (Tassa et al., 2018). While the RL community typically views state-based control as a less interesting problem, proprioceptive feedback plays a critical role in coordinated biological motor control: its disruption impairs interjoint coordination and locomotor patterning (Sainburg et al., 1995; Akay et al., 2014). The vision modality uses a third-person camera, introducing the problem of inferring body position and movement from images. We therefore use proprioception for the full suite, then compare it with image-only observations on Walker Stand, Walk, and Run across all three backbones to measure the effects of proprioception and vision on intrinsic motivation. DMC pretraining uses 1M environment steps.

Open-ended exploration We use Craftax Classic, the JAX reimplementation of Crafter, as a complementary test of open-ended exploration from vision without strong dynamical structure to exploit (unlike DMC) (Matthews et al., 2024; Hafner, 2022). Its discrete actions express abstract operations such as gathering and crafting, while achievements (which constitute reward) vary from independent goals, like waking up, to those with compositional structure, like crafting a stone pickax. Progress therefore requires discovering resources and learning how to combine them over extended behavioral sequences. We pretrain for 5M environment steps.

Multi-task transfer Walker-Stand, Walker-Walk, and Walker-Run share the same embodiment and environment dynamics with a natural compositional relationship, in that postural control supports walking and coordinated locomotion supports running. We use Stand as the source environment, then evaluate the shared model on all three rewards by training task-specific heads in imagination. For intrinsically motivated agents, "Stand source" identifies the environment in which data are collected, not the task driving exploration. The comparison with a Stand-supervised agent tests whether intrinsic exploration produces a more reusable model than experience selected for one task.

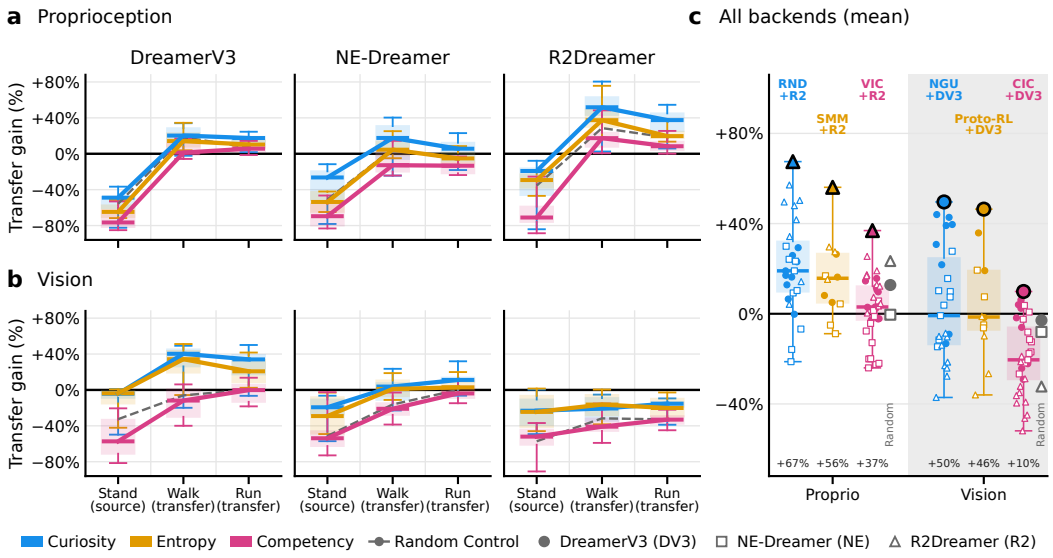

Figure 4: **Walker zero-shot transfer.** Performance on Stand, Walk, and Run after 1M pretraining steps in the Stand environment from (a) proprioception and (b) vision. Transfer gain is return minus the same-backend, same-modality Stand-supervised reference, divided by task-supervised DreamerV3 return on the evaluation task. Boxes span each family's interquartile range across methods, with medians and full-range whiskers. (c) Each method/backend pair's gain averaged over Walk and Run; labels identify family maxima within each modality.

For DreamerV3, we additonally test multitask transfer in RoboKitchen, a Franka Panda arm manipulation experiment in a fixed workspace with multiple distinct goals (Mendonca et al., 2021). Here, the source task is moving the kettle and the transfer tasks are interacting with a light-switch and sliding cabinet (Appendix Figure 11).

Few-shot adaptation After intrinsic pretraining, we reinitialize the task actor, critic, and reward predictor, keep the world model and replay, and train the fresh heads for 5k offline updates with the model fixed. We then unfreeze the model and collect task-directed experience—200K steps for Walker Stand-to-Run and 500K for Craftax—under the new task reward head, whose gradients now flow through the world model. These budgets are 20% and 10% of the corresponding pretraining budgets, which also also corresponds to the amount of steps the task-supervised agents needs to reach near-optimal performance. Further, we compare how the adaptation performance scales with offline updates (5K, 50K, and 100K optimizer steps) using the learned simulator only versus online environment interaction.

3 Results

3.1 Zero-shot Generalization

Intrinsically motivated experience supports task control without task-directed collection, but the choice of intrinsic objective matters (Figure 3a). Curiosity methods form the strongest overall group on the DMC suite, with Disagreement achieving 61.4% of the task-supervised IQM, compared with 21.1% for a random control and the family median being 50.1% over 35.5% for entropy and 25.7% for competency (Figure 3c). However, no intrinsic method can quite catch up to the task-supervised reference on the suite aggregate. Performance on individual environments is shown in Appendix Figure 8.

Craftax exposes a stronger dependence on the world-model representation. With DreamerV3, Vanilla Curiosity achieves the highest intrinsic score while Disagreement falls below Random Control despite leading the DMC suite (Figure 3b). Thus, the best DMC PROPIO drive does not transfer

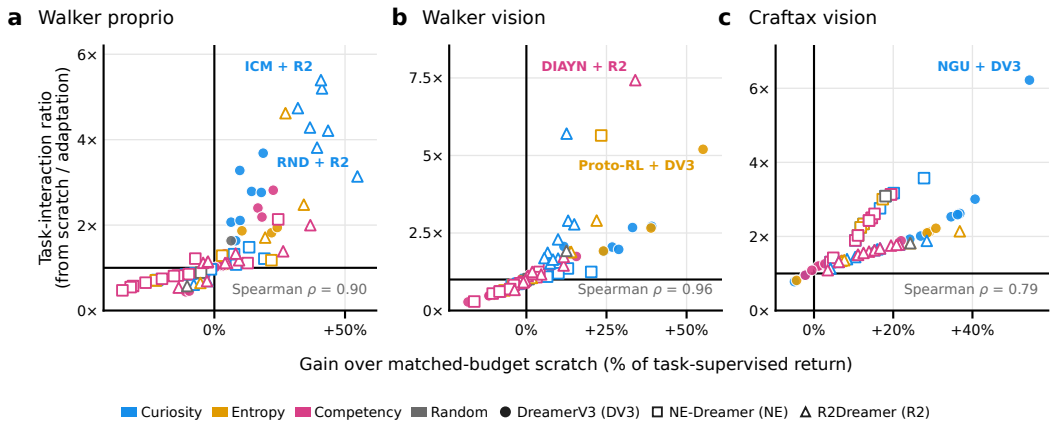

Figure 5: **Adaptation gain and sample efficiency.** All three backbones after 200K Stand-to-Run adaptation steps with (a) proprioception or (b) vision, and (c) 500K Craftax steps. Horizontal position is gain over a from-scratch agent with the same backend, modality, and task-step budget, normalized by task-supervised return. Vertical position is the ratio of from-scratch to adaptation steps at matched return (Appendix D). Labels identify the largest efficiency ratio in each panel. These ratios exclude pretraining and offline updates. Spearman correlations are descriptive across method/backend pairs.

to Craftax. The family summaries reveal an interesting interaction where curiosity's clear advantage with DreamerV3 narrows or reverses in the reconstruction-free models (Figure 3 d).

The walker transfer experiment tests the extent to which intrinsic motivation builds a model useful for zero-shot generalization to multiple distinct reward functions. In Figure 4, zero marks the performance supported by Stand-supervised experience, rather than training separately for each target task. The largest mean Walk/Run gains are RND with R2-Dreamer for proprioception (+67% of the task-supervised reference, Figure 4 a) and NGU with DreamerV3 for vision (+50% Figure 4 b). SMM with R2-Dreamer and Proto-RL with DreamerV3 also yield strong gains (+56% and +46%, respectively).

Our comparison between modalities also qualifies a universal preferences for distinct world-model representations across algorithm families. R2-Dreamer can support strong transfer from proprioception, while DreamerV3 pairs particularly well with curiosity and entropy under vision (Figure 4 c). For example, DreamerV3's curiosity median on Walk rises from +20% with proprioception to +40% with vision, whereas R2-Dreamer's falls from +52% to -21%. Richer observations therefore do not uniformly improve intrinsic pretraining, instead their value depends on how the world model represents them. Competency methods occasionally transfer well, but their family distributions generally show lower gains.

3.2 Few-shot Adaptation

In these experiments, we reinitializing the task heads us to test whether pretraining transfers useful world-model parameters and replay without inheriting the pretrained task policy. Figure 5 compares two measures of adaptation performance: points to the right of zero achieve higher return than training from scratch at the same task-step budget, while points above $1\times$ require fewer task interactions to reach a matched return. On Craftax, NGU with DreamerV3 attains the strongest result, reaching a final score of 15.8 versus 5.9 from scratch after 500K steps (a 54.4% normalized gain). Figure 6 makes this matching more explicit; here, horizontal lines locate adapted agents' returns on the learning curve of the supervised agent training from scratch. For NGU, this corresponds to approximately 3.11M steps, a $6.22\times$ improvement in sample complexity.

The best pairing differs on Walker. ICM with R2-Dreamer yields the largest task-interaction ratio from proprioception ($5.39\times$), while DIAYN with R2-Dreamer leads the vision comparison ($7.43\times$). Thus, the strongest individual adaptation result need not belong to the family with the strongest overall performance. Nor is the best zero-shot representation necessarily the best starting point

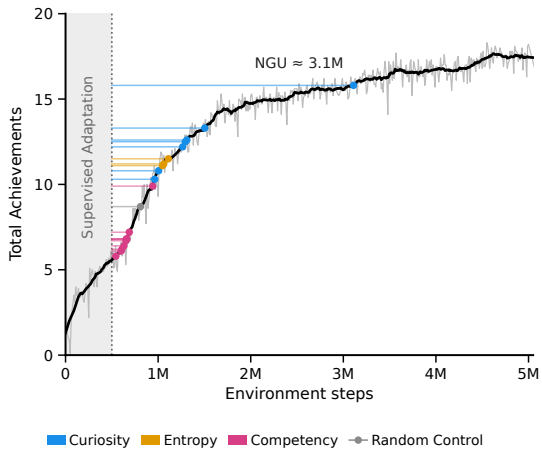

Figure 6: **Craftax equivalent from-scratch steps.** Horizontal lines match DreamerV3 agents' returns after 500K adaptation steps to the scratch curve (gray: raw; black: 150K-step smoothing).

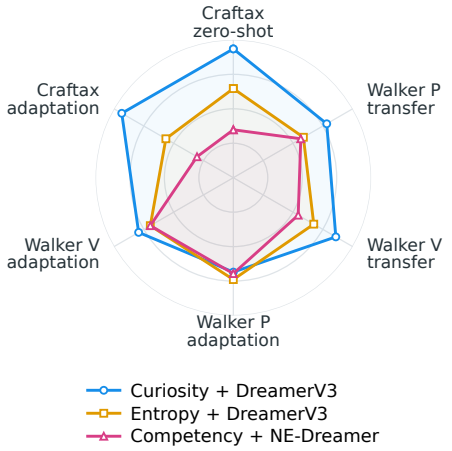

Figure 7: **Selected family profiles.** Top-four mean rankings scaled (0–100) using each family's highest-mean backend across six settings. P/V denote proprioception/vision. (Appendix D.3).

for supervised adaptation–after task reward and task-directed data update the model, a different representation may become advantageous.

Return gain and sample efficiency are positively associated but need not rank methods in absolute terms. Rather, one may be willing to sacrifice performance if environment interaction is expensive, preferring a method with higher sample-complexity. For example, ICM with DreamerV3 has the highest Walker proprio endpoint gain, but RND with R2-Dreamer has the best Walker Proprio sample complexity (9 a). In contrast, Proto-RL with DreamerV3 has the largest Walker-vision endpoint gain (+55.1%), whereas DIAYN with R2-Dreamer has the largest efficiency ratio. Reaching a reference score quickly and achieving the best final return are distinct benefits; we report both. Notably, NGU significantly outperforms all other algorithms in Craftax Vision, suggesting this environment benefits from an interaction between episodic and lifelong novelty due to the nature of its achievement structure.

3.3 World-model Representations and Family Performance

Figure 7 compares family performance across six shared settings. Selecting each family's highest-mean-rank backend for its top four methods gives DreamerV3 for curiosity and entropy, and NE-Dreamer for competency. Each curve uses the same backend and four methods on every spoke. Curiosity has the strongest mean profile, while entropy leads proprioceptive adaptation. Thus, broad usefulness does not guarantee the best method in every setting. Indeed, the highest mean-ranked individual method is Proto-RL on DreamerV3 and SMM on R2-Dreamer, both in the entropy family. Across the full comparison, curiosity has the highest family mean on every backbone, whether using the top four or all family members.

The profile shapes also show why zero-shot and adaptation results should be distinguished: relative strength at one stage need not persist when task reward begins to update the model. Ranks are computed separately within each backend, so selecting a backend here identifies a family's strongest relative profile, not its highest absolute return. Absolute comparisons come from Figures 3–5. reconstruction is particularly effective for visual zero-shot control, while reconstruction-free models support strong proprioceptive transfer and can benefit substantially from task-directed adaptation.

3.4 Intrinsic Estimators, Algorithm Complexity, and Experience

DADS versus rDADS DADS illustrates why estimator choice matters. It predicts the next world-model state from the current state and skill, rewarding transitions more likely under that skill than

alternatives (Sharma et al., 2020). Our reverse estimator, rDADS, asks how much the transition improves skill inference beyond the current state, replacing a high-dimensional next-state density problem with two low-dimensional skill-density problems. It improves every Walker zero-shot Walk/Run comparison across three backends and two modalities (Tables 5–7). Across nine backend/task/modality adaptation settings, its mean gain is +12.9% versus DADS's +1.2% (Appendix Figure 10). Yet it wins five of six Walker adaptation settings and loses all three Craftax comparisons. Skill prediction may be easier in some representations, but conditioning and density-estimator differences prevent attributing these gains to prediction direction alone.

When do bells and whistles help? In the same nine-setting comparison (Appendix Figure 10), Vanilla Curiosity (+17.7%) exceeds ICM (+12.8%) and C-TeC (+0.8%), while Disagreement (+19.8%) exceeds the implemented MAX (+18.0%) and CIG (+10.7%). Added feature learning or more elaborate uncertainty estimates therefore do not guarantee better world models. Yet episodic novelty in NGU improves RND's mean gain from +20.1% to +23.8%, and Proto-RL's learned prototype features improve on APT (-1.7% to +25.8%). These results support retaining simple latent-space baselines while adding additional mechanistc structure, such as explicit interactions between short and long-term memory.

More imagination or more experience? Our offline training procedure fits the reward predictor to replay and trains the policy in imagination, with a fixed world model and no new interaction. Tenfold fresh-head training, from 5K to 50K updates, improves initial control in both environments (Appendix Figure 9). On Walker, mean normalized gain after 200K task steps rises from +35.2% to +49.1%. On Craftax, initial gain rises from +21.4% to +30.8%, but gain after 500K steps falls from +37.5% to +35.1%. Better offline control therefore need not improve online adaptation. Both stages use imagination, but online adaptation adds task-directed data and model updates. Thus, these results suggest adaptation cannot benefit from a misaligned world model, even if it supports better initial control.

4 Discussion

What makes experience valuable before an agent has a particular task to solve? An unfamiliar experience may have little immediate use, yet teach the learner something about its world that becomes indispensable when its goals change. This reality is fundamental to the role of intrinsic motivation as a foundation for world-model learning. Our study supports the feasibility of building reusable world models through a broad range of intrinsic drives, and offers a common framework for asking what makes those drives effective. The quality of the model an agent builds provides a window into the value of its self-directed experience.

Curiosity has the strongest overall family performance, but no intrinsic drive is best across environments, modalities, and representations. Reconstruction supports strong visual zero-shot control, while reconstruction-free models can excel in proprioceptive transfer and adaptation. This dependence favors designing intrinsic motivation and world-model representations jointly. Further, simple intrinsic rewards remain competitive with more sophisticated, computationally expensive algorithms. Finally, more offline training with a fixed model can improve initial control without improving later adaptation after interacting with the environment. Together, our results demonstrate the rich interplay between intrinsic motivation and world-model building, providing an avenue for future work studying how these two mechanisms should be co-designed.

Scope and limitations Our experiments assess world models through downstream control, without directly measuring predictive fidelity or establishing that model building is the biological purpose of intrinsic motivation. Training-seed coverage is generally limited to three due to the sheer diversity of algorithms, backends, and experiment combinations. Representation strategies and intrinsic estimators differ in several implementation choices, so these are not isolated causal ablations. Future work could explore additional environments and modalities, as well as alternative world model formulations beyond Dreamer.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation and execution, including implementation of aspects of the project using Codex. Additionally, we used generative AI tools to aid and polish writing, retrieve and discover related work, and create and format plots with ChatGPT and Claude. Codex-assisted implementation work was independently reviewed by the primary author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code for our experiments is available at as part of the supplementary material. A public repository will be released should the paper be accepted.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018. URL https://arxiv.org/abs/1807.10299.

Turgay Akay, Warren G. Tourtellotte, Silvia Arber, and Thomas M. Jessell. Degradation of mouse locomotor pattern in the absence of proprioceptive sensory feedback. *Proceedings of the National Academy of Sciences*, 111(47):16877–16882, 2014. doi: 10.1073/pnas.1419045111. URL https://doi.org/10.1073/pnas.1419045111.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.06038.

Andrew G. Barto. Intrinsic motivation and reinforcement learning. In Gianluca Baldassarre and Marco Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47. Springer, 2013. URL https://people.cs.umass.edu/~barto/IMCleVer-chapter-totypeset2.pdf.

George Bredis, Nikita Balagansky, Daniil Gavrilov, and Ruslan Rakhimov. Next embedding prediction makes world models stronger. *arXiv preprint arXiv:2603.02765*, 2026. URL https://arxiv.org/abs/2603.02765.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.12894.

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*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://openreview.net/forum?id=Skc-Fo4Yg.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 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*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

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/2605.20878v1.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5306–5315, 2020. URL https://proceedings.mlr.press/v119/kim20e.html.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/debf482a7dbdc401f9052dbe15702837-Abstract.html.

Misha 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*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/091d584fced301b442654dd8c23b3fc9-Abstract-round2.html.

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

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/99bf3d153d4bf67d640051a1af322505-Abstract.html.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, and Jakob Nicolaus Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35104–35137, 2024. URL https://proceedings.mlr.press/v235/matthews24a.html.

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*, volume 34, 2021. URL https://arxiv.org/abs/2110.09514.

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*, 2026. URL https://arxiv.org/abs/2603.02008.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1509.08731.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-Dreamer: Redundancy-reduced world models without decoders or augmentation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Je2QqXrcQq.

Aran Nayebi. What capable agents must know: Selection theorems for robust decision-making under uncertainty. *arXiv preprint arXiv:2603.02491*, 2026. URL https://arxiv.org/abs/2603.02491.

Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007. doi: 10.1109/TEVC.2006.890271. URL https://www.pyoudeyer.com/ims.pdf.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071, 2019. URL https://proceedings.mlr.press/v97/pathak19a.html.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687, 2025. URL https://proceedings.mlr.press/v267/richens25a.html.

Robert L. Sainburg, M. Felice Ghilardi, Howard Poizner, and Claude Ghez. Control of limb dynamics in normal subjects and patients without proprioception. *Journal of Neurophysiology*, 73(2): 820–835, 1995. doi: 10.1152/jn.1995.73.2.820. URL https://doi.org/10.1152/jn.1995.73.2.820.

Jürgen Schmidhuber. Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pp. 1458–1463. IEEE, 1991. doi: 10.1109/IJCNN.1991.170605. URL https://doi.org/10.1109/IJCNN.1991.170605.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9443–9454, 2021. URL https://proceedings.mlr.press/v139/seo21a.html.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1907.01657.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5779–5788, 2019. URL https://proceedings.mlr.press/v97/shyam19a.html.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*, 2018. URL https://arxiv.org/abs/1801.00690.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11920–11931, 2021. URL https://proceedings.mlr.press/v139/yarats21a.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Ruihan Zhao, Kevin Lu, Pieter Abbeel, and Stas Tiomkin. Efficient empowerment estimation for unsupervised stabilization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.07356.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL fly? analysis and ingredients for mutual information skill learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xoIeVdFO7U.