

Toward Curiosity-Driven Embodied Learning Through World Models

Alyn Kirsch Tornell

CMU-RI-TR-26-84

July 13th, 2026

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:

Prof. Aran Nayebi, Chair
Prof. Yonatan Bisk
Reece Keller

*Thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Robotics*

© Alyn Kirsch Tornell

CMU-RI-TR-26-84, 2026

Abstract

Curiosity allows animals and humans to learn through interaction without explicit instruction. Rather than chasing an external reward, a curious agent builds a model of its world and explores through it. This thesis asks how principles of natural curiosity can be translated into embodied agents, and what world models are needed as bodies, action spaces, and environments become more complex. We study this progression across three settings of increasing embodied complexity in simulation, using animal behavior and neural dynamics as both inspiration and evaluation targets.

We first study futility-induced passivity in larval zebrafish, in which an animal whose swimming no longer moves the visual world becomes passive before trying again, a transition carried by a neural–glial circuit that accumulates evidence of futility. We introduce 3M-Progress, a model-memory objective that compares an online world model with a frozen memory of normal action consequences, and train an embodied virtual zebrafish using intrinsic motivation alone. The agent reproduces the active–passive cycling observed in the animal more faithfully than standard curiosity objectives, and its latent dynamics recapitulate the fast neuronal and slow glial timescales of the biological circuit.

We then extend this perspective to walking *Drosophila*, whose articulated body and richer action repertoire pose a more complex embodied setting, where the question shifts to which action is being prepared. A simulated fly, controlled by a recurrent high-level agent over a fixed low-level central pattern generator, walks on a spherical treadmill that mirrors experiments in which a premotor region predicts a spontaneous turn seconds before it begins. The environment and pipeline are complete, and the remaining experiment tests whether the agent’s recurrent activity anticipates future turn direction before movement.

Finally, we ask whether this approach extends to larger, more open-ended environments closer to those in which human curiosity operates, using a social navigation task from ego-centric vision as the intended pretraining stage for a harder object-interaction setting. Preliminary results reveal important limitations: a world-model agent does not learn even the navigation task that a model-free agent solves, improving at predicting its observations while its behavior does not, a signature of objective mismatch.

Together, these projects treat world models as the substrate through which curiosity is expressed and assessed. Purpose-built models can reproduce biologically observed behavior and its internal dynamics in a controlled setting, while extending curiosity to richer settings

depends on whether the underlying agent can first learn the environment reliably: a result in the zebrafish, an in-progress extension in the fly, and a diagnostic limit in the open-ended case.

Table of Contents

List of Figures	v
1 Introduction	1
2 Intrinsic Goals in Virtual Zebrafish	4
2.1 Introduction	4
2.2 Background	6
2.3 Approach	10
2.4 Results	15
2.5 Discussion	18
3 Extending Curiosity-Driven Embodied Learning to <i>Drosophila</i>	22
3.1 Introduction	22
3.2 Background	23
3.3 Approach	25
3.4 Results	30
3.5 Discussion	31
4 Scaling Curiosity-Driven Embodied Learning to Open-Ended Environments	32
4.1 Introduction	32
4.2 Background	32

4.3	Approach	34
4.4	Results	35
4.5	Discussion	36
5	Conclusion	38
	Bibliography	41
	Appendix Additional Details for Intrinsic Goals in Virtual Zebrafish	44

List of Figures

Figure 2.1	The zebrafish futility-to-passivity circuit	7
Figure 2.2	The virtual zebrafish agent and environment	10
Figure 2.3	Agent architecture and 3M-Progress	12
Figure 2.4	Behavioral alignment	16
Figure 2.5	Model-brain alignment	17
Figure 2.6	Latent dynamics and neural-glial timescales	21
Figure 3.1	The IPS pre-turn signal	24
Figure 3.2	The virtual <i>Drosophila</i> platform	27
Figure 3.3	Hierarchical fly agent architecture	28
Figure 3.4	The controlled fly comparison pipeline	29
Figure 4.1	Objective mismatch in the open-ended setting	36

Chapter 1

Introduction

A child left alone with a new set of toys, given no task and no reward, will reach for them, turn them over, and test what happens, exploring simply because exploration is intrinsically valuable. This kind of self-directed behavior is not unique to humans. Animals explore their environments, act without explicit instruction, and adapt when the consequences of their actions change, all without the externally specified rewards on which artificial agents are usually trained. This capacity for curiosity, for learning through interaction rather than through supervision, is what this thesis sets out to understand and to build. It matters both for neuroscience, where it helps explain natural behavior, and for artificial intelligence, where it points toward agents that learn robustly without dense supervision.

Reinforcement learning gives a way to make this idea precise. There, curiosity becomes intrinsic motivation: an agent generates its own reward from interaction rather than receiving it from a task, so that it can learn without external supervision. Yet the exploration it produces often looks nothing like an animal's: an agent rewarded for surprise can be captured by a screen of static, which it can never learn to predict and so never stops finding surprising, exploring it endlessly while learning nothing. A world model is only a starting point; what makes exploration meaningful is how curiosity is defined over that model, and whether the model is good enough to define it over in the first place.

Animals suggest what that structure might be. They do not treat all novelty alike; they organize behavior around internal models of how their actions affect the world, judging from experience what is worth doing next rather than reacting to raw surprise. A world model is the natural substrate for this, because predicting how the world responds to an agent's actions is what lets the agent evaluate them. World models enter this thesis in two connected ways. In the intrinsic rewards we study, curiosity is computed from a world model, from the gap between what an agent predicts its actions will do and what actually follows; and in the final setting we ask whether the same idea can be carried to world-model agents that learn to act inside a model of their own.

This thesis studies curiosity expressed through world models across three embodied settings of increasing complexity, using animal behavior and neural dynamics as both inspiration and evaluation. The guiding idea is that producing animal-like behavior calls for curiosity grounded in embodied interaction and shaped by prior experience, rather than generic novelty seeking. We study three questions in turn: in larval zebrafish, whether continued action remains worthwhile; in walking *Drosophila*, which action is being prepared; and in large-scale open-ended environments, whether a world model can be learned reliably enough to support curiosity at all.

The first project studies futility-induced passivity in larval zebrafish. When a fish’s swimming no longer produces the expected visual motion, it first increases its effort, then stops and becomes passive, and after a period begins swimming again, cycling between active and passive states. The behavior is a clean probe of curiosity because it is driven not by any external reward but by the animal’s own assessment of whether its actions are effective. Building on whole-brain recordings of this behavior, we ask whether an embodied agent trained on intrinsic motivation alone can reproduce both the behavior and the neural–glial dynamics that accompany it. The method we introduce, Model-Memory-Mismatch Progress (3M-Progress), compares an online world model of the current environment against a frozen memory of normal action consequences, and rewards progress in that mismatch rather than raw prediction error. The agent reproduces the active–passive cycle more faithfully than standard curiosity objectives, and its internal dynamics recapitulate the fast neuronal and slow glial timescales of the biological circuit. This is the main completed result of the thesis. The second project extends the framework to walking *Drosophila*, whose articulated body and richer behavioral repertoire pose a more complex embodied setting. Here the question shifts from whether to keep acting to which action is being prepared. Whole-brain imaging of walking flies has found that a premotor region, the Inferior Posterior Slope (IPS), carries a left-right asymmetry that predicts the direction of a spontaneous turn seconds before it begins. We build a simulated fly, controlled by a recurrent high-level agent over a fixed low-level central pattern generator, that walks on a spherical treadmill mirroring the biological preparation, together with a pipeline that aligns the agent’s internal activity to its turns. The environment and analysis are complete; the remaining experiment tests whether an agent selected for behavior develops recurrent activity that anticipates future turn direction, as the IPS does. The fly is therefore an in-progress extension rather than a second completed result.

The third project asks whether this approach reaches the larger, visually rich environments closer to those in which human curiosity operates. Because a curiosity signal defined over a world model is only as good as the world model beneath it, we first ask a more basic

question: whether a world-model agent can learn such an environment at all, using a social navigation task from egocentric vision as the pretraining stage of an intended navigation-then-rearrangement pipeline. It reveals a limitation. The world-model agent does not learn the navigation task even though a model-free agent solves it, improving steadily at predicting its observations while its behavior does not, a signature of objective mismatch. The limitation is specific and informative: in a visually rich environment the challenge lies at the world-model substrate itself, and a model that can solve the task has to come before curiosity can be added on top.

Read together, the three projects make a single argument. A world-model-based intrinsic objective can reproduce not only biologically observed behavior but the internal dynamics beneath it, when the environment is controlled enough for the objective to take hold; carrying the same idea into richer, open-ended settings depends first on whether the agent can build a reliable model of its world at all.

Chapter 2

Intrinsic Goals in Virtual Zebrafish

2.1 Introduction

Animals do not only act in response to external rewards or immediate physiological needs. They also engage in self-directed behaviors: exploring, adapting, and changing their actions even when no explicit task objective is present. This capacity for autonomy is central to natural intelligence, allowing organisms to learn from their own interactions with the world and adjust their behavior when those interactions no longer produce the expected consequences. In artificial agents, however, autonomous behavior remains difficult to reproduce. Standard reinforcement learning agents are typically trained to optimize predefined rewards, while many intrinsic motivation methods rely on generic signals such as novelty, prediction error, or uncertainty. These signals can encourage exploration, but they do not necessarily capture what makes exploration useful for an embodied animal: the relationship between the agent’s own actions and their effects on the world.

This limitation becomes especially important in open-ended environments, where goals are not fixed and the consequences of action can change over time. An agent driven only by prediction error may become attracted to inherently unpredictable stimuli, such as the classic “noisy TV” problem, rather than learning stable and meaningful interactions with its environment. Similarly, an agent driven only by novelty may continue seeking new states without developing structured behavioral transitions. Animals appear to solve this problem differently. Their exploratory behavior is not simply a response to novelty or surprise, but is shaped by whether their actions remain effective. When an animal’s own actions stop producing the expected sensory consequences, this mismatch can drive a change in behavior. This suggests that useful intrinsic motivation should be grounded in embodied interaction: agents should care not only about what is surprising, but about whether their own actions continue to matter.

In this chapter, we study this question through a specific biological case: futility-induced passivity in larval zebrafish. In this behavior, zebrafish initially swim in response to visual feedback, but eventually become passive when their swim commands no longer produce the expected visual consequences [1]. This behavior provides a natural example of reward-free autonomy because the animal is not optimizing an explicit external reward. Instead, it appears to regulate its behavior based on the mismatch between its internal expectations and the outcomes of its own actions. This makes futility-induced passivity a useful test case for asking how an intrinsic objective can produce structured, animal-like behavior in an embodied agent.

We present Model-Memory-Mismatch Progress, or 3M-Progress, as a biologically motivated intrinsic objective for autonomous behavior. The objective rewards an agent for tracking the divergence between an online world model, which changes with current experience, and a fixed prior world model distilled from ethological experience. Rather than rewarding raw prediction error, 3M-Progress focuses on how the agent’s current model of the world changes relative to prior expectations. This allows the agent to respond when self-generated actions stop producing their expected sensory consequences, while avoiding the instability of objectives based only on novelty or surprise. We instantiate this objective in an embodied zebrafish agent and train it without access to behavioral or neural data from the animal.

The resulting agent reproduces key features of zebrafish futility-induced passivity. Trained only with this self-supervised intrinsic objective, the agent develops active–passive behavioral cycling similar to that observed in larval zebrafish. More importantly, the model also predicts whole-brain neural-glial dynamics under a strict unit-to-unit mapping [2]. This alignment is important because the model is not fit directly to neural activity. Instead, the neural match emerges from an objective chosen to produce robust self-supervised behavior. This makes the model more than a descriptive fit to the data: it offers a normative account of why this behavior may arise and how neural-glial dynamics may support autonomous action selection. This chapter presents the zebrafish project as a single computational argument. We begin with the biological phenomenon of futility-induced passivity, then describe the intrinsic motivation objective it motivates, the embodied agent used to test it, and the behavioral and neural-glial alignment that results. Together, these results suggest that intrinsic motivation should be grounded in embodied world models and in the changing relationship between an agent’s actions and their consequences. This provides the central evidence for the thesis: animal-like autonomy can emerge when agents are motivated not by generic novelty, but by progress in understanding how their own actions affect the world.

2.2 Background

Futility-Induced Passivity in Larval Zebrafish

Larval zebrafish provide a useful biological setting for studying reward-free autonomous behavior because they exhibit a repeatable transition when swim commands fail to produce the visual feedback normally associated with movement [1]. This behavior, known as futility-induced passivity, has been characterized at whole-brain, single-cell resolution, providing both a behavioral target and a neural target for evaluating a computational model of intrinsic motivation.

A larval zebrafish that repeatedly tries to swim, but whose swimming no longer changes what it sees, will eventually stop trying. This behavior is known as *futility-induced passivity*. The transition does not occur immediately. When the animal’s actions first lose their expected sensory effect, it initially increases its effort, producing stronger and larger-amplitude swim attempts as though trying to overcome the discrepancy. Only after tens of seconds of repeated failure does the animal abruptly enter a passive state and stop swimming for several to tens of seconds, before spontaneously resuming and repeating the cycle [1]. In prolonged recordings, active and passive epochs are long and well separated, with active epochs lasting roughly 36 s and passive epochs roughly 24 s on average. The behavior therefore appears as a slow alternation between active and passive states rather than as random fluctuation in swim rate.

This behavior is informative because no external reward directly explains the switch. The animal is not rewarded for becoming passive, nor punished for continuing to swim. Instead, the transition reflects an internal assessment of whether continued action is effective. For this thesis, the active-to-passive transition provides a concrete behavioral signature of motor futility: a successful model should reproduce the animal’s tendency to continue acting at first, accumulate evidence that its swim attempts are ineffective, and then switch into a passive state.

The behavior is elicited in a virtual-reality preparation in which swim commands and visual feedback can be experimentally decoupled [1]. The larva is restrained, and swim attempts are measured from motor nerve activity rather than from physical displacement. A projected visual grating provides optic flow, and the experimenter controls the *motosensory gain*, which determines whether swim attempts produce the visual feedback normally associated with forward movement.

Under *closed loop*, the gain is positive. A swim bout moves the grating backward, producing the optic flow the animal would normally generate by swimming forward against a current. In this condition, the fish swims rhythmically at about one bout per second and behaves as if it is holding station in a stream. Under *open loop*, the gain is set to zero. The fish still issues

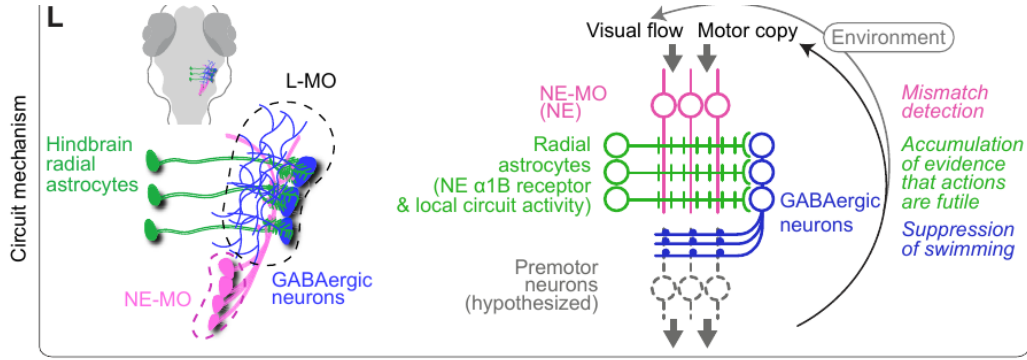


Figure 2.1. The three-stage neural–glial circuit underlying futility-induced passivity. Noradrenergic neurons in the medulla oblongata (NE-MO) detect failed swim attempts, radial astrocytes in the lateral medulla oblongata accumulate this failure signal over repeated failures, and downstream GABAergic neurons suppress swimming once the accumulated evidence is sufficient. Adapted from Mu et al. [1].

swim commands, but those commands no longer move the grating, so the expected visual consequence of action never arrives. This decoupling between motor command and sensory feedback drives passivity. A key control replays, in open loop, the same visual motion that the fish generated during closed-loop swimming. The animal still becomes passive, even though the visual input is matched to successful swimming. This shows that passivity is not caused by the visual scene alone, but by the loss of contingency between action and sensory outcome. Additional controls rule out fatigue, physical struggling, and a simple drive to maintain position, supporting the interpretation that the relevant variable is the mismatch between action and expected consequence.

Whole-brain calcium imaging during the behavior, enabled by the optical transparency of the larva and light-sheet microscopy [3], resolves the active-to-passive transition into three computational stages carried by distinct cell populations (Figure **Figure 2.1**). First, noradrenergic neurons in the medulla oblongata (NE-MO) detect failed actions. These neurons respond weakly when a swim produces the expected visual feedback and strongly when it does not. By randomizing the motosensory gain bout by bout, Mu et al. [1] show that these neurons encode the magnitude of sensory-motor mismatch, increasing with motor vigor when visual feedback is absent and with the degree of conflicting visual feedback.

Second, radial astrocytes in the lateral medulla oblongata (L-MO) integrate this failure signal over time. These cells receive noradrenergic input through the $\alpha 1B$ adrenergic receptor, and their intracellular calcium ramps upward over successive failed swim attempts. The ramp rises fastest when feedback is always withheld and does not rise when swimming remains effective. This graded dependence on feedback probability is consistent with evidence

accumulation: the circuit does not react to a single failed action, but integrates repeated evidence that action is no longer producing its expected effect.

Third, once the accumulated astrocytic signal is sufficiently high, downstream GABAergic neurons suppress swimming and induce passivity. The causal evidence supports this circuit interpretation. Ablating L-MO radial astrocytes reduces passive behavior, while chemogenetic or optogenetic activation of astrocytic calcium is sufficient to induce passivity; ablating NE-MO similarly disrupts the behavior [1]. The circuit therefore implements a concrete computation: detect failed action, accumulate evidence over time, and switch behavioral state once that evidence crosses a threshold. The fact that the slow integrative component is carried by glia is biologically important, but for the computational argument of this thesis the key point is that the circuit identifies the form of the behavior to be modeled.

Intrinsic Motivation and Curiosity Objectives

In artificial agents, the problem of acting without an external reward is usually studied through *intrinsic motivation*: reward signals generated by the agent itself rather than supplied by an external task. Intrinsic motivation includes several families of objectives. Some methods reward novelty or state coverage, encouraging the agent to visit unfamiliar parts of the environment. Others reward competence or skill discovery, encouraging the agent to acquire diverse and controllable behaviors. This chapter focuses on a third family: curiosity objectives based on world models, where experiences are valuable when they improve or challenge the agent’s predictive understanding of the world.

Within this world-model-based family, the ideal signal is *prediction gain*: the improvement in a world model caused by a new observation. If an experience makes the model better at predicting future data, then that experience is treated as useful. However, prediction gain is difficult to compute directly because it requires evaluating the model before and after learning on a validation distribution that is usually unknown. Practical curiosity methods therefore use tractable approximations to this ideal, including prediction error, novelty-based proxies, disagreement, and learning progress. Formal definitions and implementation-level details for these intrinsic motivation baselines are provided in Appendix A.

One common approximation is *prediction error*. Methods such as the Intrinsic Curiosity Module (ICM) reward the agent when its forward model is surprised by what happens next, often in a learned feature space [4]. This objective is simple and effective in many sparse-reward settings: states with high prediction loss are treated as intrinsically rewarding. Its weakness is that high error does not necessarily mean useful learning. A state can be unpredictable because it is not yet learned, but it can also be unpredictable because it is inherently random or uncontrollable. This leads to the classic “noisy-TV” failure mode, in which an agent continues to seek stimuli that remain surprising but cannot be mastered [5].

A second approximation rewards *novelty*. Count-based exploration, pseudo-counts, and Random Network Distillation (RND) encourage the agent to visit states or transitions that are rare under its past experience [6]. These methods provide a coverage signal: go where the agent has not been before. This is useful when exploration is the main difficulty, but novelty is not the same as learnability or control. A state can remain novel because it is stochastic, irrelevant, or disconnected from the agent’s actions. For this reason, novelty-based rewards can reinforce behavior that does not improve the agent’s world model in a meaningful way. A third approximation is *disagreement*. Disagreement-based exploration rewards the variance across an ensemble of predictive models [7]. If the models agree about what will happen next, the state is treated as familiar; if they disagree, the state is treated as informative. This resembles query-by-committee in active learning, where the learner seeks data that would resolve uncertainty among competing hypotheses. Disagreement can separate some forms of uncertainty from raw prediction error, but it still depends on the behavior of a moving set of learned predictors. In complex or stochastic environments, ensemble members may continue to disagree, and the method also requires maintaining multiple models.

Progress-based objectives reward improvement rather than error itself. Instead of asking whether the model is currently wrong, they ask whether it is becoming less wrong. This distinction is important because irreducible noise can keep prediction error high, but it should not produce sustained learning progress. γ -Progress makes this idea tractable by comparing the current world model to an exponential moving average of its past predictions or parameters [8]. This provides a constant-memory approximation to prediction gain, rewarding changes in model performance over time rather than the absolute level of prediction error. These baselines test different hypotheses about what should make an experience intrinsically rewarding: surprise, novelty, uncertainty, or improvement. They are therefore the natural computational points of comparison for the model introduced in this chapter and for the later embodied-agent experiments in this thesis. However, the zebrafish behavior suggests a more specific requirement. The animal does not simply respond to unfamiliar or surprising visual input; it changes behavior when its own swim attempts stop producing their expected sensory consequences. This motivates an intrinsic objective based on motor futility: whether self-generated actions continue to have the effects predicted by the agent’s world model. The next section introduces Model-Memory-Mismatch Progress, which builds on progress-based curiosity but anchors the reward to a fixed reference model in order to produce stable, repeatable behavior.

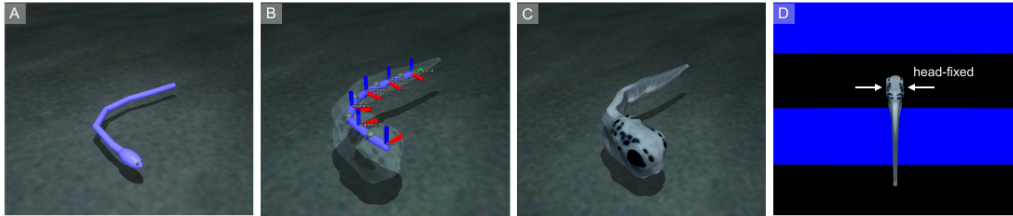


Figure 2.2. The embodied virtual zebrafish. (A) A six-link swimmer body in a fluid environment with passive and active currents. (B) The agent actuates joint torques to swim and navigate. (C) A cosmetic skin matching the appearance of a larval zebrafish. (D) The head-fixed open-loop environment, built to match the experimental parameters of Mu et al. [1]. From Keller et al. [2].

2.3 Approach

We introduce an embodied zebrafish agent trained only through intrinsic motivation in a physically simulated environment. The agent is not trained to imitate zebrafish behavior, and its internal activity is not fit to neural recordings. Instead, it learns from the sensory consequences of its own actions. After training, we compare the agent’s behavior to futility-induced passivity in larval zebrafish and compare the activity of its internal modules to the neural-glial circuit described above. This setup allows the model to be evaluated as more than a descriptive fit: if an intrinsic objective chosen for self-supervised embodied behavior reproduces both the animal’s behavior and the structure of its neural dynamics, then the objective provides a candidate computational account of how the behavior could arise.

The approach has four parts. First, we construct a realistic virtual zebrafish body whose actions generate sensory feedback through simulated physics. Second, we define two environments: an ethological environment in which swimming produces normal visual consequences, and an open-loop environment in which swim commands no longer move the visual scene. Third, we train a recurrent agent whose policy is driven only by intrinsic reward computed from learned world models. Finally, we introduce Model-Memory-Mismatch Progress, or 3M-Progress, which rewards changes in the mismatch between a frozen ethological memory and an online model of the current environment.

Embodied Zebrafish Agent

The agent is designed around the sensorimotor loop measured in the zebrafish experiment. Its motor commands move a physical body, the body interacts with a fluid environment, and these interactions determine the visual feedback available to the agent. This makes the embodiment and environment part of the learning problem rather than implementation

details: the intrinsic objective is evaluated on the sensory consequences of movement, just as the biological experiment manipulates the visual consequences of swimming.

The virtual zebrafish is implemented in the MuJoCo physics engine [9] using the n -link swimmer and inertial fluid model from the DeepMind Control Suite [10]. The body is actuated through joint torques and moves through a simulated fluid that exerts both passive and active forces. The agent receives egocentric visual input and proprioceptive state, allowing it to learn from the same broad categories of information available to a swimming animal: what it sees, what it does, and how its body changes. Its visual input is set at a resolution comparable to the visual acuity of larval zebrafish [11], so that the sensory consequences of swimming are constrained by a biologically plausible perceptual channel.

This embodied setup is important for the thesis because curiosity is not treated as a generic drive toward novel states. It is evaluated in a body that must act through continuous motor commands and in an environment where those actions produce structured sensory consequences. The model therefore follows the broader program of using realistic virtual animals to study behavior and neural computation [12–15], but differs in that the behavior is not produced by imitation learning. The agent learns its policy from an intrinsic objective.

Ethological and Open-Loop Environments

The agent is trained and evaluated in two environments designed to reproduce the key sensorimotor contrast in the zebrafish experiment [2]. In the *ethological environment*, the agent swims freely against a current. Its swim commands move its body and produce visual feedback, giving it experience with the normal relationship between motor output and self-generated optic flow. A forward world model trained in this environment becomes the agent’s *ethological memory*: a fixed prior over how actions usually change the sensory stream under naturalistic swimming conditions.

In the *open-loop environment*, the agent is head-fixed and its swim commands no longer move the visual scene. This environment is constructed to match the relevant experimental parameters of the zebrafish setup, including the visual grating, scale, contrast, and motion statistics. It recreates the key manipulation from the biological experiment: the agent can still issue swim commands, but those commands no longer produce the expected visual consequences. This is the simulated analogue of motor futility.

The two environments are used in different stages of learning. The ethological memory is learned first and then frozen. During open-loop training, a separate online world model is learned from the current environment. When the agent is passive, the transition dynamics of the ethological and open-loop environments are more similar, so the two models tend to agree. When the agent attempts to swim, the models diverge: the ethological memory predicts sensory consequences of self-motion, whereas the open-loop environment produces

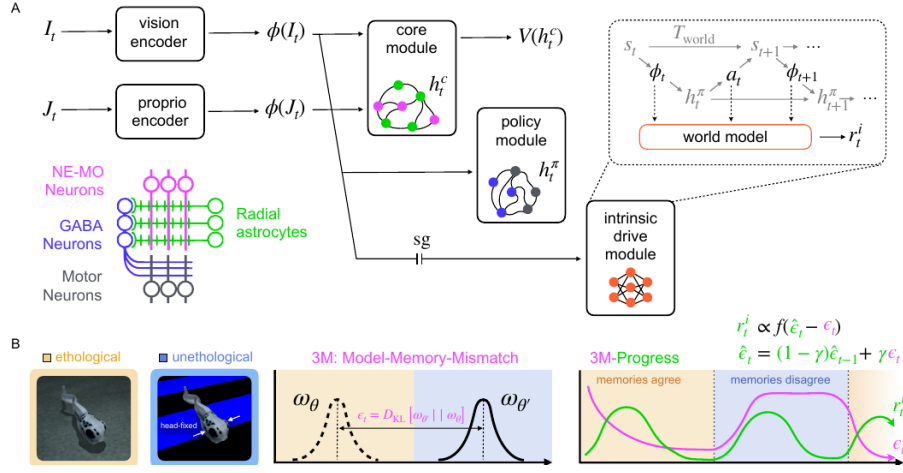


Figure 2.3. Agent architecture and 3M-Progress. **A.** Egocentric visual input I_t is encoded by a small residual network ϕ_I , while proprioceptive state observations J_t are encoded by a small multi-layer perceptron ϕ_J with shortcut paths to the core and policy modules. Sensory features are passed to recurrent LSTM core and policy modules, h_t^c and h_t^π , which learn a state-value function and stochastic policy, respectively. The intrinsic drive module is a small multi-layer perceptron that parameterizes a forward dynamics model over sensory features from an environment with transition dynamics T_{world} . **B.** 3M-Progress compares two memories learned from environments with different transition dynamics: a frozen ethological memory, or prior ω_θ and an online world model $\omega_{\theta'}$. Their divergence defines the model-memory mismatch, which is passed through a leaky integrator $\hat{\epsilon}$ to generate the intrinsic reward r_t^i . From Keller et al. [2].

little or no corresponding visual feedback. This design makes the mismatch specific to failed self-generated action rather than to the visual stimulus alone.

Agent Architecture

The zebrafish agent is equipped with a recurrent sensory-cognitive architecture designed to support perception and action in continuous, high-dimensional environments (Figure [Figure 2.3A](#)). Because futility-induced passivity in larval zebrafish is driven by the relationship between swim commands and visual feedback, the agent’s sensory stream is centered on egocentric vision. Images are encoded by a visual encoder operating at a resolution comparable to the visual acuity of larval zebrafish [11]. The resulting sensory features, together with proprioceptive information, are passed to three modular networks: a core module, a policy module, and an intrinsic drive module.

The core and policy modules are implemented as an actor-critic architecture using Long Short-Term Memory (LSTM) networks [16] followed by feedforward decoders. The policy

outputs joint torques that control the swimmer body, and the full actor-critic system is trained end-to-end with Proximal Policy Optimization (PPO) [17] and generalized advantage estimation [18]. Although PPO is a model-free reinforcement learning algorithm, the intrinsic drive module learns a forward world model through online experience. This world model approximates the transition dynamics of the environment in sensory feature space, predicting future sensory features from the current sensory features and action. Importantly, the world model is used only to generate intrinsic reward. It is not used for planning, imagined rollouts, or model-based control.

Model-Memory-Mismatch Progress

3M-Progress compares two world models: a frozen ethological memory ω_θ and an online world model $\omega_{\theta'}$ learned in the current environment (Figure [Figure 2.3B](#)) [2]. The ethological memory encodes the sensory consequences of action under normal swimming conditions. The online model tracks the sensory consequences of action in the current environment. The instantaneous *model-memory mismatch* is defined as the Kullback-Leibler divergence between the predictions of these two models:

$$\epsilon_t := D_{\text{KL}} [\omega_\theta(\phi(s_{t+1}) \mid \phi(s_t), a_t) \parallel \omega_{\theta'}(\phi(s_{t+1}) \mid \phi(s_t), a_t)], \quad (2.1)$$

where $\phi(\cdot)$ is the sensory feature embedding and a_t is the action.

The mismatch ϵ_t is small when the current environment behaves like the ethological memory and large when it does not. Because the Kullback-Leibler divergence is non-negative, it coarsely partitions the state-action space into regions of model-memory agreement, where $\epsilon_t \approx 0$, and regions of model-memory disagreement, where $\epsilon_t \gg 0$. In the open-loop environment, passive states tend to fall closer to the agreement partition because little self-generated sensory change is expected or produced. Swim attempts tend to fall into the disagreement partition because the ethological memory predicts sensory consequences of self-motion while the open-loop environment does not provide them. Thus, ϵ_t acts as a computational analogue of a failed-action signal.

The instantaneous mismatch is then passed through an exponential filter:

$$\hat{\epsilon}_t = (1 - \gamma)\hat{\epsilon}_{t-1} + \gamma\epsilon_t, \quad (2.2)$$

where γ determines the filter timescale. This filtered mismatch provides a moving baseline over recent model-memory mismatch. Unlike standard learning-progress objectives, which compare an online model to a slowly changing memory of itself, 3M-Progress filters the mismatch between a fixed ethological memory and the online model. The result is a derivative-like signal over mismatch while preserving a stationary reference point.

The intrinsic reward is computed from the deviation between instantaneous mismatch and the filtered baseline:

$$r_t^i = |\hat{\epsilon}_t - \epsilon_t|. \quad (2.3)$$

The absolute value makes the reward symmetric: it reinforces deviations from the moving baseline in either direction. As a result, the agent is not rewarded simply for remaining in either the agreement partition or the disagreement partition. Instead, the reward encourages periodic exploration between partitions. In the zebrafish-inspired open-loop environment, this means that the agent can be rewarded both when it moves into active behavior that reveals model-memory mismatch and when it returns toward passive behavior where the two models agree more closely.

All experiments and baseline algorithms include a small action penalty,

$$r_t^a = -\lambda \|a_t\|_2^2, \quad (2.4)$$

which encourages exploration of passive behavior rather than forcing continuous movement. The key design choice is that the ethological memory is fixed. In curiosity, disagreement, and standard learning-progress objectives, intrinsic reward is coupled to a moving world model. As the model learns, prediction error can flatten, causing the reward landscape to collapse and preventing the objective from selecting a stable exploratory policy. In 3M-Progress, the online model may adapt to the current environment, but the ethological memory remains anchored to the pretraining environment. The mismatch therefore does not vanish simply because the online model improves. Instead, as the online model stabilizes, the signal-to-noise ratio of the mismatch can increase, producing more robust behavioral patterns. This fixed reference allows the agent to settle into a stable, repeatable behavioral cycle rather than losing its intrinsic reward as learning converges.

3M-Progress as a Normative Model of Neural-Glial Computation

The components of 3M-Progress also define a candidate computational mapping onto the neural-glial circuit described above. The instantaneous model-memory mismatch ϵ_t has the functional role of a failed-action signal: it becomes large when the agent attempts to swim but the expected sensory consequence does not occur. This is functionally analogous to the NE-MO neurons in zebrafish, which respond to failed swim attempts and encode mismatch between motor output and visual feedback [1].

The filtered mismatch $\hat{\epsilon}_t$ has the functional role of an accumulated failure signal. Because it is computed by an exponential filter over recent mismatch, it is analogous to radial astrocytes that accumulate NE-MO input over repeated failed swim attempts and decay during passivity. In this sense, 3M-Progress is not only an intrinsic reward objective; it also specifies

a computational decomposition that can be compared to the biological circuit.

This mapping should be understood as a functional correspondence rather than a claim of direct mechanistic identity. The model proposes that a fixed ethological memory, an online estimate of current transition dynamics, and a leaky accumulation of their mismatch are sufficient to reproduce the behavioral structure of futility-induced passivity and to generate internal variables comparable to the neural and glial signals observed in larval zebrafish.

2.4 Results

We evaluate whether 3M-Progress reproduces futility-induced passivity at two levels. First, we ask whether the trained agent develops the same active–passive behavioral cycling observed in larval zebrafish. Second, we ask whether the agent’s internal dynamics align with the neural and glial activity recorded during the behavior. In the main text, we focus on comparisons against the intrinsic motivation baselines introduced above: ICM, RND, Disagreement, and γ -Progress. Additional details on these baselines, along with the full comparison including model-free and data-derived controls from Keller et al. [2], are provided in Appendix A. Across these comparisons, the central question is whether 3M-Progress produces stable, animal-like autonomous behavior by comparing current experience to a fixed ethological memory.

Behavioral Alignment

Behavioral alignment is measured by comparing swim-power dynamics in the animal and the agent. In the zebrafish, swim power is computed from motor nerve activity; in the agent, it is computed from the norm of joint torques [2]. These quantities are not physically identical, but they provide comparable time series of motor output. The relevant comparison is therefore not the absolute scale of movement, but the temporal structure of behavior around active and passive transitions.

Agents trained with 3M-Progress develop active–passive cycling within roughly ten million environment steps and maintain this behavior over training (Figure Figure 2.4A). This is a stronger criterion than matching average activity: a successful model must reproduce the alternation between vigorous swimming and passive pauses. 3M-Progress achieves the highest behavioral alignment of all tested objectives, approaching the inter-animal ceiling defined by the similarity between zebrafish individuals (Figure Figure 2.4B). By contrast, the curiosity baselines show transient strategies early in training but do not maintain the stable transition structure of the animal. Some collapse toward passivity, while others continue producing activity without the long-timescale active–passive cycling characteristic of futility-induced passivity (Figure Figure 2.4A,C).

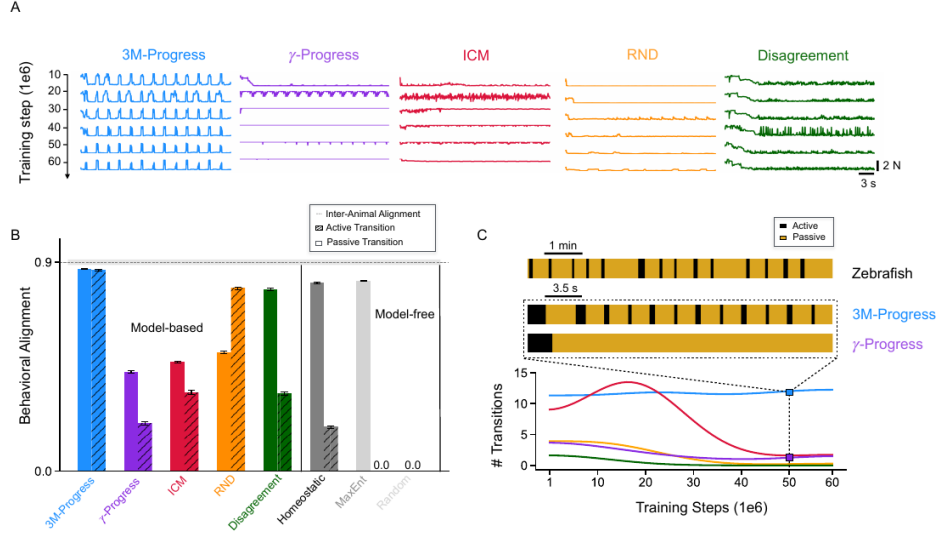


Figure 2.4. Behavioral alignment between zebrafish and intrinsically motivated agents. **A**. Swim-power traces over training for agents with different intrinsic drives. 3M-Progress develops sustained active-passive cycling, while the other objectives either lose transitions or collapse to less structured behavior. **B**. Pearson correlation between agent and zebrafish swim power around active and passive transitions. 3M-Progress approaches the inter-animal alignment ceiling and exceeds all baselines. **C**. Number of behavioral transitions over training. 3M-Progress maintains transition structure similar to zebrafish, whereas other drives lose transitions over training. From Keller et al. [2].

These results show that stable animal-like behavior does not follow automatically from giving an agent a world model or a novelty signal. The structure of the intrinsic objective matters. In this environment, 3M-Progress is the only tested objective that turns the loss of expected sensory feedback into a repeatable behavioral cycle resembling the zebrafish.

Neural-Glial Alignment

We next compare the agent’s internal activity to whole-brain calcium imaging recorded during futility-induced passivity. The biological dataset contains approximately 300,000 recorded cells, including roughly 130,000 neurons and 130,000 astrocytes [1, 2]. Both biological activity and model activations are aligned to active and passive behavioral transitions using event-triggered averaging. Alignment is then measured using a one-to-one mapping in which each biological unit is matched to the single most correlated artificial unit.

This evaluation is intentionally restrictive. Unlike linear-regression mappings that can combine many model units to predict one neural response, the one-to-one mapping asks whether individual artificial units have dynamics that directly resemble individual biological units. This strict comparison is appropriate here because the zebrafish recordings are highly reliable

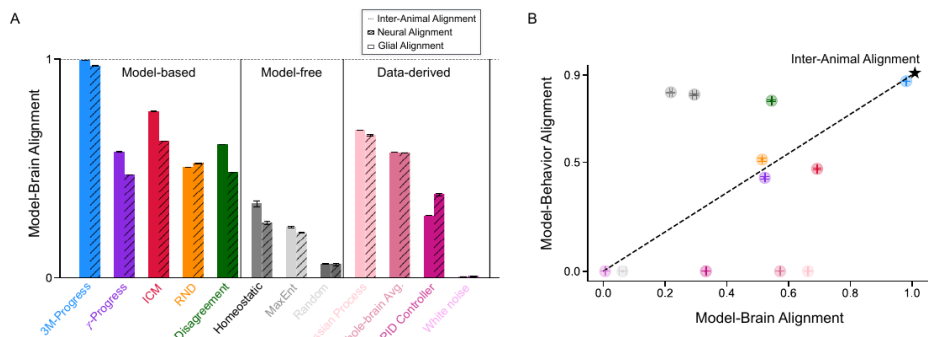


Figure 2.5. Model-brain alignment between zebrafish neural-glial activity and agent internal dynamics. **A.** Noise-corrected correlation between biological units and artificial units for model-based, model-free, and data-derived candidates. 3M-Progress achieves the strongest alignment for both neural and glial activity. **B.** Joint behavioral and neural alignment. 3M-Progress is the only model near the inter-animal ceiling on both axes. From Keller et al. [2].

across animals, producing a near-unity inter-animal ceiling against which model alignment can be evaluated.

Under this criterion, 3M-Progress achieves the strongest neural-glial alignment among the intrinsic motivation objectives (Figure Figure 2.5A). It explains substantially more of the neural and astrocytic response structure than ICM, RND, Disagreement, and γ -Progress. The full comparison in Appendix A includes additional model-free and data-derived controls from Keller et al. [2], showing that the alignment is not explained by generic activity patterns, simple dynamical fits, or behavioral matching alone.

Critically, 3M-Progress is the only model that lies near the inter-animal ceiling on both behavioral and neural-glial alignment simultaneously (Figure Figure 2.5B). This joint alignment matters because either axis alone would be insufficient: a model could reproduce the behavioral transitions through an unrelated control strategy, or it could contain internal variables that resemble the neural data without producing the correct behavior. By the standard of the NeuroAI Turing Test, which requires models to match both behavior and internal function rather than either alone [19], 3M-Progress passes on this dataset.

Neural-Glial Timescales in the Agent’s Latent Dynamics

The alignment above is measured unit by unit, but the biological circuit is defined equally by its *timescales*: a fast neuronal response to each failed swim attempt and a slow glial accumulation of that failure signal [1]. To ask whether the trained agent recovers this temporal organization, we apply Principal Component Analysis to the activity of the agent’s core module around active and passive transitions [2] (Figure Figure 2.6). The dominant

latent components separate into two timescales. One component changes rapidly at the transition, tracking the onset of model-memory mismatch and resembling the transient response of the NE-MO neurons that detect failed actions. A second component evolves more slowly, ramping up as failed swim attempts accumulate and remaining elevated during passivity, resembling the integrative response of radial astrocytes. A complementary fast–slow separation appears when the agent returns from passivity to active swimming.

These latent dynamics are not fit to the recordings. The agent is never trained on neural or glial data, and the timescales emerge only from optimizing the intrinsic objective. The fast component follows directly from the instantaneous mismatch ϵ_t , and the slow component from its leaky accumulation $\hat{\epsilon}_t$, so the two timescales built into 3M-Progress reappear in the learned representation. This shows that 3M-Progress reproduces not only *that* the animal switches between active and passive states, but the distinct neural-like and glial-like timescales that underlie the switch.

Together, these results show that 3M-Progress reproduces both the active–passive behavioral cycle and the neural-glial dynamics of futility-induced passivity. The key difference from ICM, RND, Disagreement, and γ -Progress is the fixed ethological memory: failed swim attempts remain meaningful because they are evaluated against a prior learned from normal swimming, rather than against a predictor that is continually changing with the current environment.

2.5 Discussion

The zebrafish results show that an intrinsic objective can reproduce both reward-free behavior and neural-glial dynamics when it is grounded in embodied interaction. 3M-Progress provides a computational account of motor futility: the agent detects when self-generated actions stop producing their expected sensory consequences, accumulates that mismatch over time, and transitions into passivity. Because the model is trained only through self-supervised intrinsic reward and is compared to behavior and neural activity only after training, the alignment supports a normative interpretation of the biological circuit.

In this interpretation, the zebrafish circuit implements a useful computation for autonomy. NE-MO neurons provide a fast failed-action signal, radial astrocytes accumulate that signal over time, and downstream inhibitory neurons suppress swimming once accumulated evidence is sufficient. 3M-Progress reproduces the same functional structure with an instantaneous model-memory mismatch, a leaky accumulation of mismatch, and an intrinsic reward that regulates transitions between active and passive behavior. The result does not imply that the artificial units are biophysically identical to the biological cells, but it does

suggest that the neural-glial circuit may implement a computational strategy for deciding when continued action is no longer useful.

This project anchors the thesis because it links the behavioral, computational, and neural levels in a single system. The behavior is reward-free and experimentally controlled; the dataset includes whole-brain neural and glial recordings; the agent is embodied in a physically simulated environment; and the objective is trained without behavioral or neural supervision. Matching behavior alone would be weaker, because the agent could find a different control strategy. Matching neural activity alone would also be weaker, because the model might not produce the behavior. The joint alignment is what makes the result central to the thesis.

Limitations

Several limitations bound the interpretation of these results. First, the virtual zebrafish is a simplified approximation of the animal. The simulation captures a continuous swimmer body, fluid interaction, egocentric visual input, and an open-loop manipulation inspired by the biological experiment, but it does not model the detailed biomechanics, musculature, spinal circuitry, or full fluid dynamics of a larval zebrafish. The agent’s prior experience is also not identical to the animal’s developmental or homeostatic history. It learns an ethological memory in a swimming environment, but this should be understood as a computational approximation to prior experience rather than a literal model of the animal’s life history.

Second, the neural-glial correspondence is functional rather than mechanistic. The recurrent units and intrinsic-drive variables recover temporal profiles comparable to neural and astrocytic activity, but the model does not simulate noradrenergic receptor dynamics, intracellular calcium signaling, radial-astrocyte morphology, or synaptic connectivity. The model therefore makes claims about computation: what signals are useful, how they are accumulated, and how they regulate behavior. It does not claim to reproduce the biophysical mechanism by which the zebrafish circuit implements those signals.

Third, the evidence comes from a single behavioral paradigm. Futility-induced passivity is especially useful because it isolates a clear form of motor futility, but it is still only one behavior. 3M-Progress is formulated more generally: any setting with a meaningful prior over action-dependent sensory dynamics could, in principle, define a model-memory mismatch. However, the present results validate the objective only in the zebrafish open-loop setting. Its generality to richer behavioral repertoires remains an open question.

The zebrafish project provides the clearest demonstration in this thesis that a world-model-based intrinsic objective can produce animal-like embodied behavior. However, it studies one animal, one environment, and one behavioral transition: switching between active and passive states when expected sensory feedback is removed. The rest of the thesis asks how far this framework can be pushed. The next chapter turns to *Drosophila*, a different embodiment

with a richer motor system, to examine whether curiosity-driven embodied learning through world models can support more complex behavioral organization beyond the active–passive structure studied here.

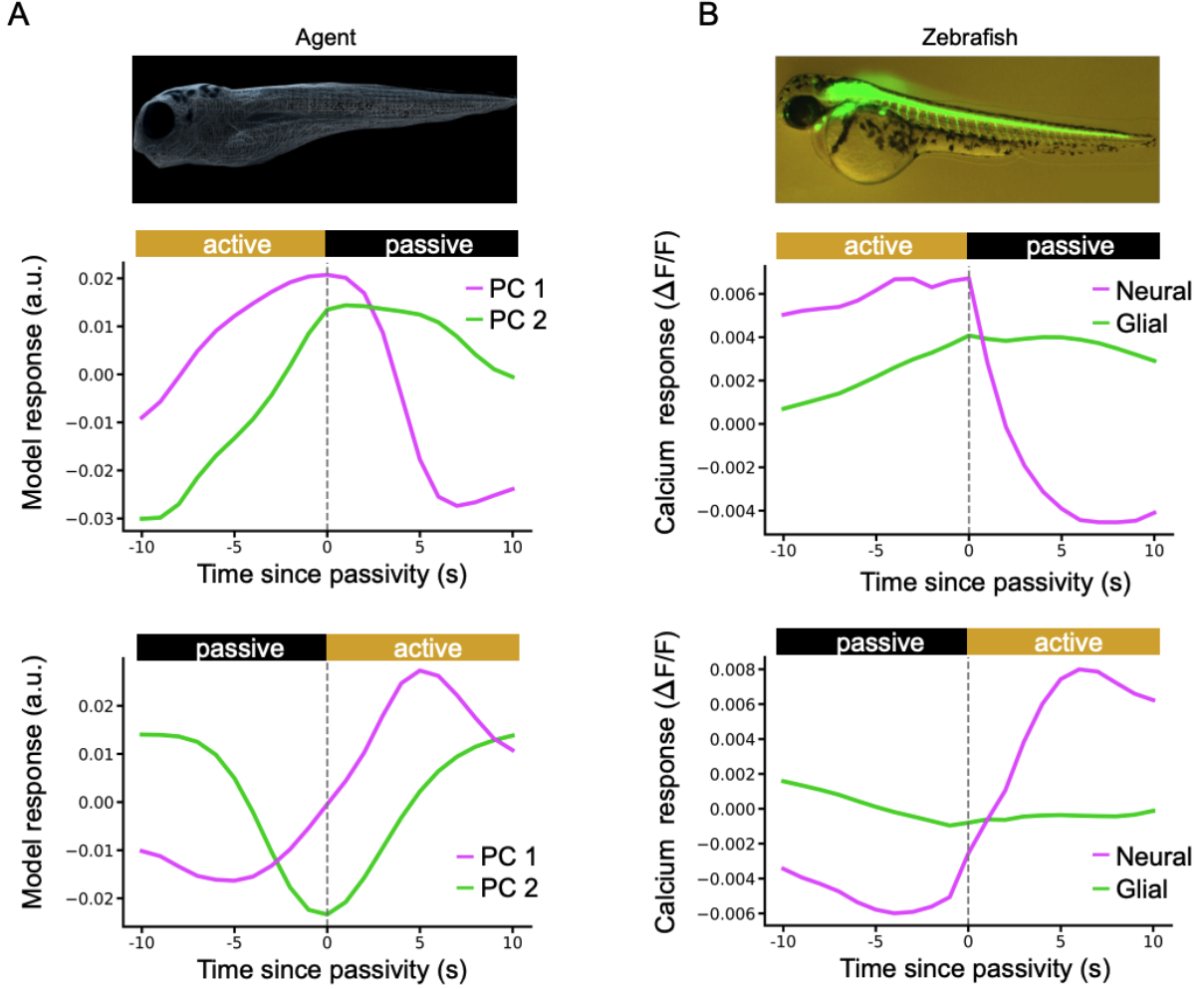


Figure 2.6. Latent dynamics of the 3M-Progress agent compared with whole-brain neural-glial activity in zebrafish, aligned to behavioral transitions (top row: active-to-passive; bottom row: passive-to-active). **(A)** The two dominant principal components of the agent’s core-module activity: PC1 (magenta) tracks a fast component and PC2 (green) a slower one. **(B)** Whole-brain neural (magenta) and glial (green) calcium responses around the same transitions in a zebrafish. The agent recovers the same fast-neural/slow-glial organization without being trained on neural data. From Keller et al. [2].

Chapter 3

Extending Curiosity-Driven Embodied Learning to *Drosophila*

3.1 Introduction

The zebrafish project provides the clearest demonstration in this thesis that a world-model-based intrinsic objective can produce animal-like embodied behavior. However, it studies one animal, one environment, and one behavioral structure: switching between active and passive states when expected sensory feedback is removed. This chapter asks how far the same framework can be pushed toward a different embodiment and a different kind of behavior.

Drosophila provides a useful next step because its body and behavioral repertoire are richer than the larval zebrafish setting studied in the previous chapter. Rather than modeling a transition between swimming and passivity, the fly project focuses on walking and turning: behaviors that require coordination across multiple legs, interaction with a physical substrate, and selection among future motor actions. This makes the fly a natural test case for extending curiosity-driven embodied learning beyond a single active–passive state transition.

The biological motivation comes from whole-brain recordings of head-fixed walking flies on a spherical treadmill. In these experiments, neural activity can be aligned to spontaneous turning behavior, allowing the analysis to ask whether the brain contains signals that predict future turns before the movement is expressed. The Inferior Posterior Slope (IPS) is especially relevant because it contains an early left-right asymmetry that predicts future turn direction several seconds before the turn. This suggests a neural target for artificial agents: if an embodied fly model develops internal variables that predict future turning behavior, those variables can be compared to the biological IPS signal.

This chapter presents the current state of the fly project. The contribution is not yet a final neural alignment result comparable to the zebrafish chapter. Instead, the focus is on building the embodied simulation and analysis pipeline needed for such a comparison. We

construct a tethered fly environment inspired by the biological spherical-treadmill setup, use a hierarchical control architecture to separate low-level leg coordination from higher-level behavioral selection, and record internal model states in a format that can be aligned to future turning behavior. The goal is to make it possible to test whether curiosity-driven embodied agents can produce internal dynamics comparable to future-action signals in a second animal embodiment.

3.2 Background

Walking *Drosophila* provide a complementary biological setting for studying how future behavior is organized during embodied locomotion. In the zebrafish project, the central behavior was a transition between active swimming and passivity when expected visual feedback was removed. In the fly, the behavior of interest is different. Rather than asking why an animal stops acting when its actions become ineffective, the fly experiment asks how neural activity is organized before a future movement is expressed.

A walking fly does not move continuously in a single smooth trajectory. Instead, it explores through locomotor bouts composed of straight or curved runs and rapid saccadic turns, separated by periods of quiescence and grooming [20]. This richer behavioral stream is useful for analysis because saccadic turns can be identified as discrete peaks in angular velocity. Although the animal’s behavior is not limited to turning, each turn has a measurable direction that can be extracted from rotational velocity, making it possible to ask whether brain activity contains information about an upcoming left or right turn before that turn is behaviorally expressed.

In the biological experiment, behavior was measured using a spherical-treadmill preparation. Head-fixed flies walked on an air-suspended foam ball in darkness while whole-brain neural activity was recorded with two-photon imaging [20]. This setting is important because the animal’s trajectory was not guided by salient spatial sensory cues. The direction of each turn therefore provides a useful behavioral target for studying internally organized action selection. During each 30-minute recording session, the ball was imaged at 50 Hz, and its motion was processed to estimate the animal’s forward and rotational velocity. Across nine flies, these recordings were registered into a common brain space, creating a dataset in which walking and turning behavior could be analyzed together with brain-wide neural activity [20].

The neural analysis takes advantage of the bilateral symmetry of the fly brain. Activity on one side of the brain was compared with activity in the corresponding region on the opposite side, making it possible to distinguish symmetric activity from left-right differences. Forward

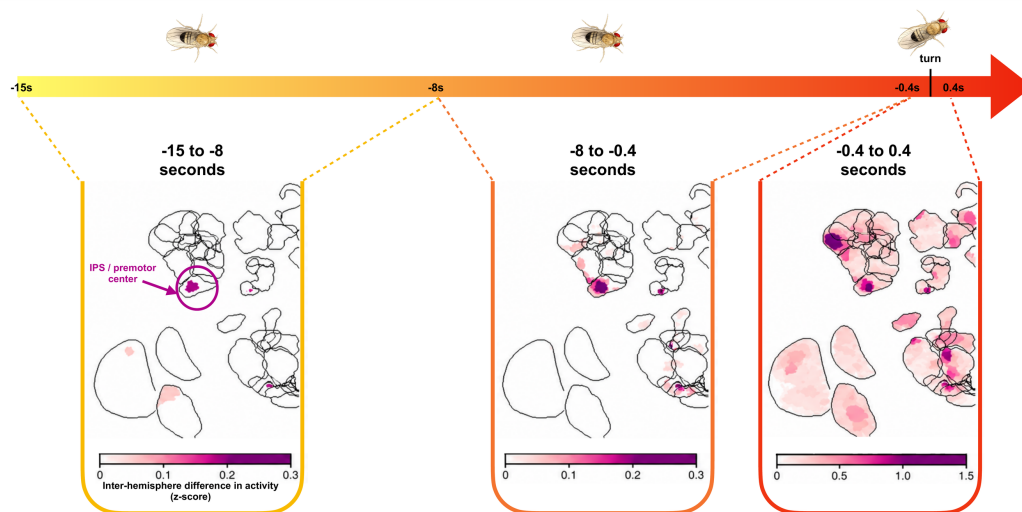


Figure 3.1. The Inferior Posterior Slope (IPS) carries a direction-specific signal seconds before a spontaneous turn. The three snapshots show the inter-hemisphere difference in neural activity at successive windows aligned to turn onset. In the -15 to -8 s window, a localized left–right asymmetry is already present in the IPS (circled) and depends on the direction of the upcoming turn; by -8 to -0.4 s it has spread to additional regions; and around the turn (-0.4 to 0.4 s) the asymmetry is brain-wide. Adapted from Brezovec et al. [20].

walking was associated with a more symmetric signal, while turning was associated with a left–right imbalance: the difference in activity between hemispheres correlated with angular velocity [20]. In other words, moving forward and turning were represented in separable neural dimensions, with turning expressed as a symmetry break between the two sides of the brain.

The important feature of this symmetry break is that it begins before the turn is executed. Turning events were identified from peaks in angular velocity and aligned to the peak of each turn, allowing neural activity to be examined in the seconds before movement. Between 15 and 8 seconds before the turn, a small premotor region centered in the Inferior Posterior Slope (IPS) already showed a left–right difference that depended on whether the animal would later turn left or right [20]. This made IPS the earliest identified source of the direction-dependent signal preceding a turn (Figure 3.1).

This early signal is important because it is not simply a reflection of the animal already turning. In the biological data, average angular velocity remained close to zero until roughly two seconds before the turn, while the IPS asymmetry was detectable earlier, in the 15-to-8 second pre-turn window [20]. The direction information was also modest rather than deterministic. IPS activity weakly but significantly predicted future turn direction across all

turns, and prediction accuracy increased when the analysis was restricted to turns preceded by stationary behavior [20]. This suggests that the IPS signal should be interpreted as a preparatory bias or future-action variable, not as a completed motor command.

The neural process therefore has a simple form: a local left-right asymmetry appears in IPS many seconds before a turn, carries information about the direction of the upcoming turn, and precedes the broader brain-wide activity associated with turn execution. For this thesis, the important computational target is not to reproduce fly volition directly, but to ask whether an embodied artificial controller can contain internal states that predict future turn direction before that direction is visible in behavior.

3.3 Approach

We introduce an embodied *Drosophila* simulation and analysis pipeline for studying future-action signals in a physically grounded walking agent. Unlike the zebrafish project, this project is not yet a completed test of an intrinsic objective against both behavior and neural activity. Instead, the goal is to build the environment, controller interface, and analysis tools needed to ask a comparable question in a richer embodiment: whether internal controller states contain information about an upcoming turn before that turn is expressed in behavior. The simulated fly is not trained to imitate fly neural activity, and its internal states are not fit to the IPS recordings. The comparison to the biological experiment is therefore functional rather than mechanistic. In the biological data, the relevant signal is a left-right asymmetry in IPS that predicts future turn direction before the turn occurs. In the model, the corresponding question is whether the high-level controller contains internal variables that can be aligned to future left and right turns and tested for information about upcoming turn direction. This does not imply that the artificial controller is biologically equivalent to IPS, but it provides a controlled setting for asking whether similar future-action variables can arise in an embodied agent.

The approach has four parts. First, we construct a tethered spherical-treadmill environment in which a simulated fly interacts with a ball through leg contact, allowing forward and rotational velocity to be extracted from the ball’s motion. Second, we use a hierarchical control structure in which a low-level controller handles leg coordination while a high-level recurrent controller selects among discrete behavioral options such as left, right, straight, stop, and groom. Third, we record behavioral variables and internal controller states across multiple rollouts from the same trained policy. Finally, we define a future-turn analysis that identifies left and right turn bouts from rotational velocity, aligns model activity to those events, and tests whether internal states predict turn direction before the turn is behaviorally

expressed.

The simulated environment is inspired by the biological spherical-treadmill preparation. The fly body is held in place while its legs interact with a ball underneath it. As the legs move, the ball rotates, providing a behavioral readout analogous to the tracked treadmill motion in the biological experiment. This setup makes it possible to estimate forward and rotational velocity from the simulated behavior, define turn bouts, and align internal model activity to those bouts.

This tethered setting is important for the same reason the virtual zebrafish embodiment was important in Chapter 2. The agent’s behavior is not generated in an abstract state space. It arises through contact between a body and an environment. Leg movements affect the ball, the ball provides the behavioral readout, and the resulting locomotor dynamics determine when the agent is walking, turning, or pausing. This makes the simulation closer to the biological setting and provides a more meaningful basis for comparing internal model activity to neural recordings.

As in the zebrafish project, the agent is trained and evaluated in two related settings. During training, the fly behaves freely in a richer environment and receives egocentric vision, odor, and proprioception,¹ moving among obstacles and three odor sources so that its actions produce structured multisensory consequences. For evaluation, the agent is placed on the tethered spherical treadmill that mirrors the biological preparation, with vision and odor removed so that it acts from proprioception alone. Its walking and turning then rotate the ball beneath it, allowing forward motion and turn direction to be recovered using the same behavioral structure as the biological experiment while matching the head-fixed, sensory-isolated condition in which the IPS signal was recorded.

The current implementation uses the NeuroMechFly body from FlyGym in the MuJoCo physics engine [9, 21, 22], a biomechanically detailed *Drosophila* model with six articulated legs and 42 leg degrees of freedom. The main technical requirement is stable contact between the legs and the spherical treadmill. The fly must be positioned so that the legs can plant on the ball without the body sinking into the surface or losing contact entirely. The environment therefore requires careful calibration of body height, ball placement, contact parameters, and default posture. Once stable contact is obtained, the environment can generate the behavioral variables needed for comparison to the biological data (Figure 3.2).

The modeling pipeline separates low-level motor execution from higher-level behavioral organization. The low-level controller is a fixed central pattern generator (CPG) that converts

¹The fly agent receives an odor channel that the zebrafish agent of Chapter 2 does not, so the input symbols are defined per chapter. In Figure 3.3, I_t , J_t , and K_t denote vision, odor, and proprioception respectively, whereas in Chapter 2 proprioception is denoted J_t .

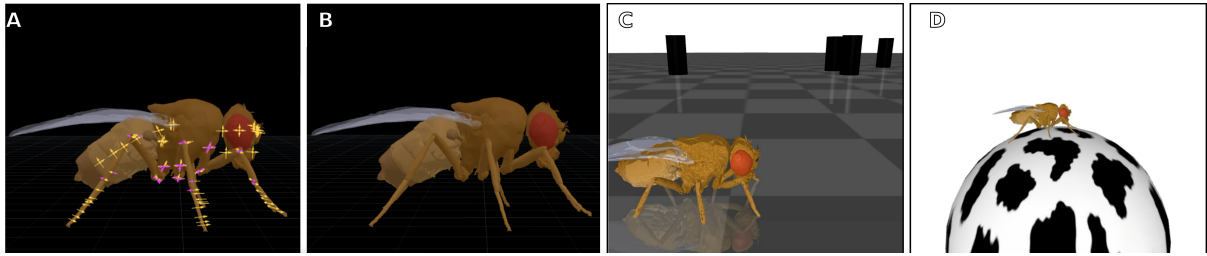


Figure 3.2. The embodied *Drosophila* platform. (A) The NeuroMechFly body with its actuated leg degrees of freedom and (B) the body at rest. (C) The training environment, in which the agent walks freely among obstacles with vision, odor, and proprioception. (D) The tethered evaluation setup, in which the head-fixed agent walks on a patterned spherical treadmill using proprioception alone, so that ball motion recovers forward and turning behavior. Built on the NeuroMechFly/FlyGym body [21, 22] in MuJoCo [9].

a discrete high-level action into the coordinated leg movements and adhesion commands required for contact-rich walking. The high-level controller is a recurrent agent that selects among discrete behavioral options (walking straight, turning left, turning right, stopping, and grooming) leaving the detailed locomotion to the fixed CPG. Following the zebrafish agent of Chapter 2, the high-level controller is a recurrent actor-critic with *separate* core and policy modules: a core module with its own LSTM that estimates the expected future return, and a policy module with its own LSTM that maintains a distinct recurrent state and selects the action. Keeping the actor and critic states separate matters here because the internal variable compared to IPS is the slowly evolving recurrent state of the policy module, not a shared value-policy representation. This hierarchy is useful because the IPS signal is unlikely to correspond directly to individual joint torques; it is more naturally compared to a higher-level internal state that evolves before a future behavioral choice is expressed. Holding the body, action space, and fixed CPG constant also allows different intrinsic objectives to be compared without changing the underlying motor system (Figure Figure 3.3).

During simulation, the pipeline records both behavior and internal model activity. Behavioral variables include forward velocity, rotational velocity, and turn bouts estimated from the movement of the ball. Internal variables include recurrent hidden states, policy-related representations, and action logits from the high-level controller. These signals are saved over time so that they can be analyzed in the same general format as the biological recordings. The central analysis aligns model activity to future turning behavior. Turn bouts are identified from rotational velocity, separated into left and right turns, and used to define pre-turn windows. The model’s internal activity can then be tested for whether it predicts future turn direction before the turn begins. This parallels the biological analysis of IPS, where

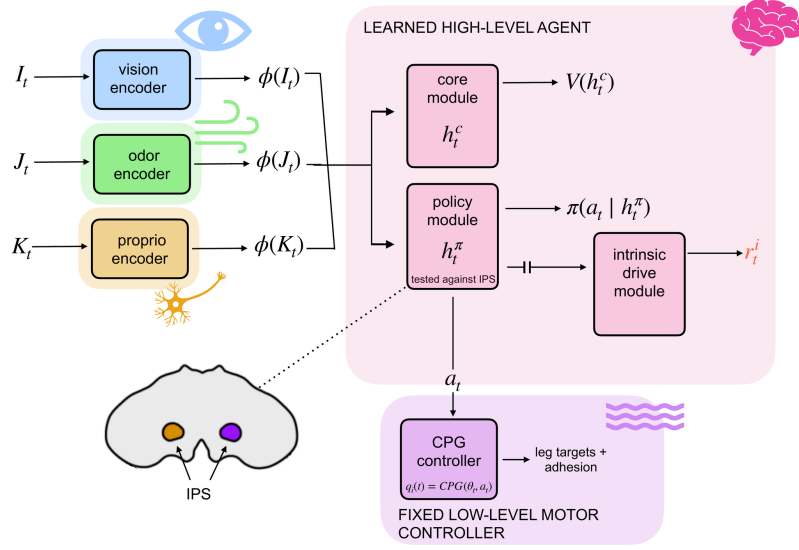


Figure 3.3. Architecture of the embodied fly agent. Egocentric vision (I_t), odor (J_t), and proprioception (K_t) are encoded and passed to a learned high-level agent with *separate* recurrent modules: a core module (h_t^c) that estimates the value $V(h_t^c)$ and a policy module (h_t^π) that produces the action distribution $\pi(a_t | h_t^\pi)$. An intrinsic-drive module supplies the reward r_t^i . The selected discrete action a_t is executed by a fixed low-level central pattern generator (CPG), which produces leg targets and adhesion. The recurrent state of the policy module is the variable compared against the biological IPS.

left-right asymmetry predicts future turning several seconds before movement.

A simple model-side analogue of the IPS signal can be constructed from the high-level controller. If the policy produces logits or probabilities over left-turn and right-turn options, then a decision variable can be defined as

$$D_m = \text{logit}_{\text{left}}(m) - \text{logit}_{\text{right}}(m), \quad (3.1)$$

where m indexes the high-level control step. Positive values indicate a bias toward left turns, while negative values indicate a bias toward right turns. If this variable predicts future turn direction before the movement begins, then it provides an artificial analogue of the IPS pre-turn asymmetry.

This variable is only one possible readout. A stronger analysis would test whether future turn direction can be decoded from recurrent hidden states or other internal representations, rather than from action logits alone. This distinction matters because action logits may reflect an imminent policy decision, while recurrent states may reveal a slower internal buildup more comparable to the IPS signal. For this reason, the planned evaluation treats neural comparison as a held-out analysis: the model should first be selected for behavior, and only

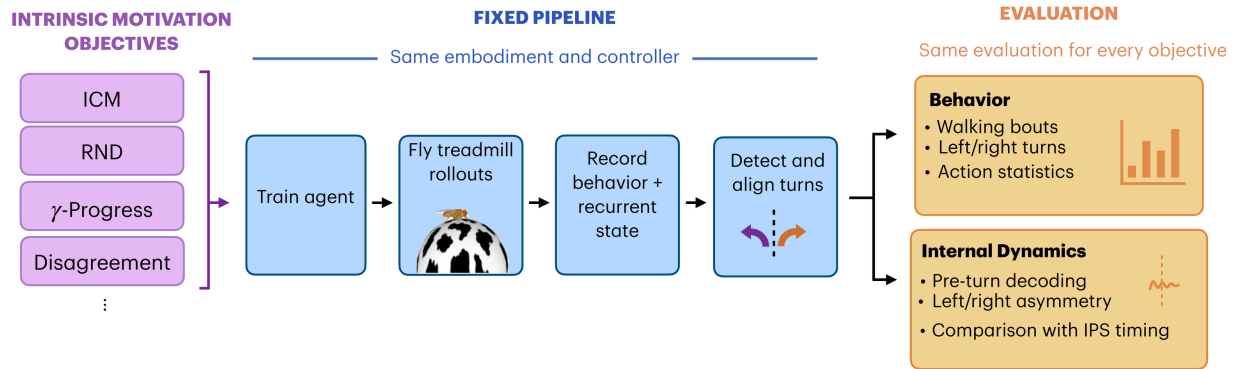


Figure 3.4. The planned comparison pipeline. The same fly body, action space, and fixed CPG controller are held constant while the high-level agent is trained under each intrinsic objective. Each trained agent is evaluated on the tethered treadmill, where behavior (walking bouts, left and right turns, and action statistics) is used for model selection. Only afterward is the held-out internal analysis applied, testing whether the policy module’s recurrent state shows pre-turn, direction-specific structure comparable to the biological IPS.

afterward should its internal activity be compared to the biological data.

As in the zebrafish project, the high-level agent is trained purely from intrinsic reward, with no external task reward for walking or turning, and its intrinsic-drive module computes reward from a learned world model over sensory features exactly as in Chapter 2. The planned experiment holds the fly body, action space, environment, and fixed CPG constant and trains the same high-level agent under a set of intrinsic objectives such as ICM, RND, γ -Progress, and Disagreement so that they can be compared on equal footing. Each trained agent is then frozen and evaluated on the tethered treadmill. Model selection is based on behavior alone: whether the agent produces stable, fly-like walking and a balanced set of left and right turns. Only afterward is the held-out internal analysis applied, testing whether the policy module’s recurrent state predicts future turn direction in the same pre-turn window as the biological IPS. Keeping the neural comparison held out matters for the same reason as in the zebrafish project: if an agent selected for behavior also develops predictive pre-turn activity, that structure was produced by the intrinsic objective rather than fitted to the brain data. This comparison across objectives is the central remaining experiment of the fly project, and its results are not yet reported here (Figure [Figure 3.4](#)).

3.4 Results

The fly project is currently less complete than the zebrafish project. The results reported here should therefore be interpreted as implementation progress rather than as a final claim of model-brain alignment. The main contribution at this stage is the construction of the tethered simulation and the analysis pipeline needed to evaluate future-turn signals in an embodied fly agent.

The current environment places a simulated fly on a spherical treadmill and allows leg contact with the ball. As the fly moves its legs, the ball rotates, making it possible to extract forward and rotational motion from the simulated setup. These behavioral variables correspond to the quantities used in the biological analysis, where ball motion is used to estimate walking and turning.

This implementation establishes the behavioral foundation for the neural comparison. Without a stable tethered environment, turn bouts cannot be defined reliably, and internal activity cannot be aligned to the same pre-turn windows used in the biological data. The environment therefore provides the necessary bridge between detailed motor control and the higher-level question of future action organization.

At this stage, the main validation criterion is behavioral. Because the low-level walking is handled by the fixed CPG, what matters is that the high-level agent generates enough left and right turns, without a strong directional bias, for turn direction to be decoded. This is the prerequisite for any meaningful comparison to IPS.

The next analysis tests whether the model contains internal variables that predict future turn direction. In the biological data, IPS activity predicts whether the fly will turn left or right during a pre-turn window roughly 15 to 8 seconds before the turn. The model-side analysis will use the same logic: identify turn bouts from simulated behavior, align internal activity to those bouts, and evaluate whether future direction can be decoded before movement begins. The planned comparison focuses on high-level recurrent dynamics rather than low-level motor commands. This is important because a signal that appears only at the moment of action execution would not be comparable to IPS. The relevant target is a slowly evolving internal variable that precedes the turn and predicts its direction before the behavioral output is visible. If such a variable emerges in an embodied agent selected for behavior, it would provide evidence that the framework can produce internal future-action signals in a second animal embodiment.

The current implementation already supports the necessary data collection. Recurrent states, policy variables, and action-related signals can be recorded over time and saved in a time-by-units format compatible with the biological analysis pipeline. This makes it possible to apply the same event-aligned decoding strategy used for the fly recordings to the artificial

agent.

3.5 Discussion

The *Drosophila* project extends the thesis beyond the specific structure of the zebrafish experiment. In zebrafish, the model addresses motor futility: whether continued action remains useful when expected sensory feedback disappears. In the fly, the target behavior is different. Walking and turning require a richer body, a different physical environment, and a different kind of behavioral organization. The project therefore tests whether the embodied, world-model-centered approach developed in the thesis can be carried into a second animal setting.

The current contribution is primarily infrastructural. The tethered spherical-treadmill environment makes it possible to generate walking and turning behavior in a setup that resembles the biological experiment. The recording pipeline makes it possible to align artificial internal states to future turn bouts and compare them to the IPS pre-turn signal. These pieces are necessary before the stronger thesis question can be answered: whether an intrinsically motivated embodied agent develops internal dynamics that anticipate future actions in a biologically meaningful way.

The project should be interpreted conservatively. Unlike the zebrafish chapter, it does not yet provide a completed behavioral and neural alignment result. The current environment must still be validated for stable locomotion, realistic turning statistics, balanced left and right turn bouts, and reliable internal readouts. Directional biases in the controller, unstable leg-ball contact, or insufficient spontaneous turning would make the IPS comparison difficult to interpret.

Finally, the connection to intrinsic motivation remains to be completed. The current pipeline establishes the embodied environment and the neural-readout analysis, but the strongest version of the project would train the high-level controller with a curiosity-driven objective rather than using only externally specified behavioral structure. In that form, the fly project would test whether world-model-based intrinsic motivation can support richer behavioral organization in a different body and motor system.

The next step is therefore clear. The environment must support stable tethered walking and spontaneous turning, the controller must generate enough behavioral diversity for analysis, and the internal dynamics must be evaluated in held-out pre-turn windows. If the model develops an IPS-like future-turn signal under those conditions, the project would extend the zebrafish result from motor futility to richer embodied behavioral organization in *Drosophila*.

Chapter 4

Scaling Curiosity-Driven Embodied Learning to Open-Ended Environments

4.1 Introduction

The zebrafish and *Drosophila* projects studied curiosity in deliberately controlled settings: a single active–passive transition driven by withheld visual feedback, and turning on a spherical treadmill. Isolating one behavior and pairing it with a biological measurement is what made those projects tractable, but it is also a step removed from the open-ended environments in which curiosity is most compelling, the large, visually rich, dynamic worlds where an agent must decide for itself what is worth exploring. This chapter takes a first step toward that regime, asking whether the world-model-based approach developed in the previous chapters can serve as a foundation for curiosity once the environment is no longer controlled.

A curiosity signal defined over a world model is only as good as the world model beneath it. Before an intrinsic objective can do anything useful, the agent has to model a rich environment well enough that the objective has real structure to act on. This chapter therefore begins not by adding a new curiosity signal but by asking a more basic question: can a world-model agent learn to solve a task in such an environment at all? The answer matters on its own, and it shows where the effort has to go before curiosity can even be evaluated in open-ended settings.

4.2 Background

The longer-term aim motivating this chapter is an agent that goes beyond its pretraining and seeks out high-quality experience for itself, by watching other agents and by experimenting with objects. This ambition is grounded in how human infants explore. Infants do not

treat all novelty alike. They distinguish agents from inanimate objects by self-initiated, biological motion, expect objects to move only when acted upon, and expect actions on objects to be efficient [23]. They infer the goals behind others’ actions and imitate the intended outcome rather than the literal behavior [24], and they preferentially attend to stimuli that are neither too simple nor too complex to learn [25]. Exploration organized around agents, objects, contact, and goals is highly structured, and this kind of agent-directed curiosity is the longer-term direction that motivated the choice of environment: Social Navigation supplies another agent to observe and follow, and Social Rearrangement supplies both objects to act on and a partner agent to coordinate with. What this chapter actually tests, however, is the prerequisite for any of this: whether a world-model agent can acquire a base model that solves the task in such an environment at all.

The environments in which human-like curiosity operates are open-ended: they contain many sources of dynamics at once, generated by objects and by other agents, with widely varying structure. A useful way to organize this variety, following the active-world-model-learning literature, is by *learnability* [8]. Some dynamics are easy to learn and quickly exhausted, such as a periodically moving object. Some are effectively unlearnable, such as the motion of wind-blown leaves, and offer no lasting structure to acquire. The interesting regime lies in between: dynamics that are complex but learnable, such as the behavior of another agent. Curiosity that is useful for building world models must allocate attention toward this complex-but-learnable regime and away from the unlearnable end of the spectrum. Failing to do so produces the *white-noise problem*, in which an agent perseverates on high-entropy but unlearnable stimuli because they remain perpetually surprising [4, 8].

World models are an appealing substrate for curiosity because they change how an agent can explore. Most curiosity signals are retrospective: a state is scored as novel or surprising only after it has been visited, so an agent driven by such a signal must first reach an informative state before it can be drawn toward it. Plan2Explore [26] was designed to overcome exactly this limitation, and it is the closest precedent for what this chapter attempts. Built on the same Dreamer world model used here, it defines its intrinsic reward as the disagreement of an ensemble of latent dynamics models, high precisely where the model is still uncertain. Because that reward lives in the model’s latent space, it can be evaluated on imagined rollouts, so the agent estimates which states would be informative and plans toward them before visiting them, with no task reward at all. Plan2Explore is therefore a direct instance of this thesis’s guiding idea, curiosity expressed through a world model, and a concrete motivation for pairing a world model with curiosity: it takes the Dreamer world model and adds an exploratory drive on top of it, so that it seeks out informative experience by anticipation rather than by chance.

The catch is that all of this presupposes a world model good enough to plan through. A world model is only a useful substrate if learning to predict translates into learning to act, and this is not guaranteed. *Objective mismatch* [27] names the gap directly: a world model is trained to predict accurately, which is not the same objective as acting well, so a model can become steadily better at prediction without the policy becoming any better at the task. The problem is acute in high-dimensional observation spaces, where a reconstruction objective is dominated by getting appearance right, much of which is irrelevant to the task. In that setting, objective mismatch is the white-noise problem reappearing at the level of the world-model objective itself: predictive progress is spent on the parts of the observation stream that are rich but task-irrelevant. This is why the present chapter asks about the substrate before it asks about curiosity: an approach like Plan2Explore can only pay off if the underlying world model converts prediction into task performance, and whether DreamerV3 does so in a photorealistic environment is precisely what is in question.

To study this question we use two related tasks from Habitat 3.0 [28]. In *Social Navigation*, an agent must find and follow a moving human while maintaining a safe distance of roughly one to two meters, acting from an egocentric RGB camera in a photorealistic indoor scene that it has not seen during training. The task is dynamic in a specific sense: the goal is a moving, reactive agent that is itself an obstacle, so the agent must anticipate where the person will go rather than home in on a fixed target. In *Social Rearrangement*, a robot and a humanoid avatar must collaborate to rearrange objects to their goal positions, so success requires not only manipulating objects and changing the state of the environment but coordinating with the other agent, for instance yielding space so the humanoid can pass. The intended design was a pretrain-then-adapt pipeline: use Social Navigation to establish a base model, and then move to Social Rearrangement to study adaptation and, ultimately, curiosity. As the candidate world-model substrate we use DreamerV3 [29], a reconstruction-based latent world model that learns its policy inside imagined rollouts of the model.

4.3 Approach

This pipeline is a concrete instance of the ordering that motivates the chapter, a base model before curiosity. Social Navigation is the pretraining stage, whose role is to establish that base model; Social Rearrangement is the stage at which adaptation and, eventually, curiosity would be evaluated.

The learnability framing of Section 4.2 explains why the world-model approach in particular should struggle: for a reconstruction-based model, a photorealistic observation stream is dominated by task-irrelevant appearance, so the model spends its predictive capacity on the

rich but irrelevant parts of the input.

Because a Dreamer-style agent learns its policy inside its own world model, a curiosity signal is only meaningful once that model can solve the task. The necessary first step is therefore to test whether the base model can learn the task at all, before any intrinsic objective is introduced.

This ordering determines the diagnostic design. We train DreamerV3 with the task reward on the Social Navigation stage and ask only whether it can learn this stage at all. A PPO agent [17] solves Social Navigation on this benchmark [28], so the task is solvable and the question is only whether a world-model agent can learn it too. If the world-model agent cannot learn the navigation stage even with a task reward available, then any failure observed after moving to rearrangement or after adding a curiosity signal would be uninterpretable: the fault could lie in the world model, the policy optimization, the partial observability, an added curiosity signal, or their interaction. Establishing whether the substrate works at all is a prerequisite for asking the curiosity question, and because the base model could not be established in the navigation stage, the study did not proceed to social rearrangement or to any intrinsic objective.

4.4 Results

This section reports the preliminary experiments and should be read conservatively. The goal is not to claim that curiosity in open-ended environments is solved or refuted, but to report what the experiments show about the difficulty of using an existing world-model agent as a substrate in this setting.

The central result is that the DreamerV3 agent does not learn Social Navigation, the pre-training stage on which the rest of the intended pipeline depended. Across roughly thirteen thousand training episodes it records zero successes, and task return fluctuates without trend. The failure occurs at the level of the substrate: the agent does not acquire successful behavior even with the task reward available and before any intrinsic objective is added, so the study never reached the harder rearrangement task. A PPO agent [17] solves Social Navigation, so the task is solvable and the reward signal is usable; the failure is therefore not intrinsic to the task but specific to the world-model approach or to its interaction with this environment.

The failure is informative because of its structure (Figure [Figure 4.1](#)). Over the same training, the world model’s reconstruction loss falls from 178 to 36 and continues to decrease: the model is demonstrably learning to predict its observations better and better. Task return, meanwhile, shows no corresponding trend. The agent becomes steadily more accurate at

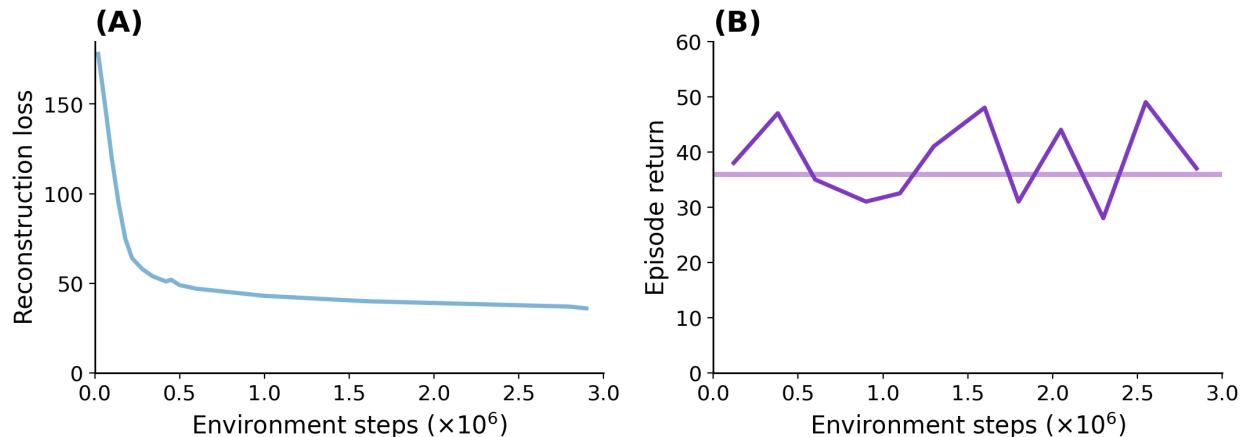


Figure 4.1. Objective mismatch on Social Navigation. **(A)** The world model’s reconstruction loss falls from 178 to 36 and continues to decrease, so the model is learning to predict its observations. **(B)** Episode return over the same training oscillates around a flat mean (horizontal line) with no upward trend, so task return does not improve. Predictive accuracy and task return are decoupled, the signature of objective mismatch.

predicting the scene without becoming any better at the task.

Taken together, these observations are a concrete signature of objective mismatch [27]. Reconstruction loss and task return move independently: the model optimizes prediction, which is not the same as optimizing task return, and the two come apart. In this high-dimensional visual setting the reconstruction objective is captured by appearance, the rich but task-irrelevant part of the observation stream, so predictive progress does not translate into task performance. This is the white-noise problem of Section 4.2, arising here in the world-model objective rather than in an explicit curiosity signal.

4.5 Discussion

The result explains a design choice in the study: no curiosity signal was ever added. In a Dreamer-style agent the policy is learned entirely inside the world model, so the quality of behavior is bounded by the quality of that model. If the base model never learns to solve the task, a curiosity objective layered on top has nothing to work with, because there is no successful behavior for it to shape. The negative result is therefore also constructive: it points to an ordering in which a base model that solves the task must come first, before any curiosity signal can help. This mirrors the trajectory of recent robotics, where a large, capable base model is pretrained and only afterward adapted to a downstream task [30]; on that view, curiosity belongs to the second stage, once such a base model exists.

What the experiment localizes is where that base model breaks down. Because a PPO agent solves Social Navigation, the failure of DreamerV3 on the same task cannot be blamed on the environment or its reward; it sits in the world-model substrate itself, in the form of objective mismatch. In a visually rich environment, a reconstruction-based model spends much of its predictive capacity on appearance, most of which is irrelevant to acting, so predicting the scene better does not make the agent act better.

These conclusions are bounded by what was actually run. The finding is a negative result for one world-model agent on one task, and further work is needed before we can say how general the failure is. We do not claim that world models cannot serve as substrates for curiosity in open-ended environments, only that this substrate, in this setting, did not learn to solve the task, which its use presupposes.

The direction that follows is to establish a base model that solves the task at all: to pretrain in Social Navigation until behavior is reliable and only then adapt, with an intrinsic objective, to the harder rearrangement task, so that curiosity can finally be assessed rather than pre-empted by substrate failure. The next chapter draws the three projects together.

Chapter 5

Conclusion

This thesis places the world model at the center of curiosity in two connected ways. The intrinsic rewards it studies are computed from a world model, from the gap between what an agent predicts its actions will do and what actually follows; and its final project carries the same idea to world-model agents, which learn to act inside a model of their own. In both, a curious agent is not chasing an external reward but building a model of its world and exploring through it, so that curiosity is a signal defined over the agent’s own predictive model of how its actions change the world. We studied this idea across three embodied settings of increasing complexity: larval zebrafish, walking *Drosophila*, and large-scale open-ended environments.

The zebrafish project provides the central result. Trained on an intrinsic objective alone, with no access to behavioral or neural data, an embodied agent using Model-Memory-Mismatch Progress reproduced futility-induced passivity: it recovered the animal’s active–passive behavioral cycle and, under a strict unit-to-unit mapping, the whole-brain neural–glial dynamics that accompany it, including the fast neuronal and slow glial timescales. Because the objective was selected for self-supervised embodied behavior and only afterward compared to the biology, the match supports a normative reading of the circuit: comparing current action–outcome dynamics against an ethological prior, and integrating the resulting mismatch over time, is a computation useful for autonomy that both the model and the animal arrive at.

The *Drosophila* project extends the framework to a richer body and a different behavior, not whether to keep acting, but which action is being prepared. We built an embodied fly on a spherical treadmill with a hierarchical controller, a recurrent actor–critic with separate core and policy modules over a fixed central pattern generator, together with a full pipeline that aligns internal activity to turns so that the pre-turn recurrent state can be compared to the biological IPS signal that predicts turn direction before movement. The platform and the analysis pipeline are complete; the comparison across intrinsic objectives is the remaining experiment, and its results are not reported here. The fly project is therefore a preliminary

extension rather than a second completed result.

The open-ended project marks a limitation of the current approach. Moving toward the visually rich, dynamic environments in which human-like exploration actually occurs, we asked whether a state-of-the-art world-model agent, DreamerV3, could serve as the substrate for curiosity, using Social Navigation as the pretraining stage of an intended navigation-then-rearrangement pipeline. Here the world-model approach reaches its limit: with the task reward alone, DreamerV3 does not learn the navigation stage, while its reconstruction loss falls steadily, a signature of objective mismatch, even though a model-free agent solves the same task. The limitation is specific and informative: for a world-model agent in a visually rich environment, the challenge lies at the substrate itself, in overcoming the reconstruction burden, before curiosity can be added.

Read together, the three projects make a single argument. A world-model-based intrinsic objective can reproduce not only biologically observed behavior but the internal dynamics underneath it, at least when the environment is controlled enough for the objective to take hold: in the zebrafish it recovers both the active-passive cycle and the neural-glial dynamics that accompany it, and the fly extends the same idea to a second embodiment. Scaling to the visually rich, open-ended settings where human-like exploration operates runs into a different and more basic obstacle. There the agent must build a reliable world model of its own before any intrinsic objective can act through it, and getting that model to work is the open problem this thesis points to. A model that solves the task has to come before curiosity can help.

Three threads follow directly. The first is completing the fly comparison: running the intrinsic objectives through the existing pipeline and testing whether an agent selected for behavior develops IPS-like pre-turn dynamics, which would extend the zebrafish result to internally generated action selection in a second embodiment. The second is the open-ended setting, where Chapter 4 traced the obstacle to objective mismatch: a reconstruction-based world model spends its capacity on appearance rather than on learning to act. Overcoming that, so that a world-model agent can learn such an environment at all, is the open problem, and how curiosity should then enter, across both pretraining and adaptation, remains to be worked out. The third is moving beyond prediction-based signals toward the agent-directed curiosity that organizes infant exploration: rewarding what an agent learns about other agents, objects, contact, and goals.

At its core, this thesis is about curiosity expressed through a world model: an agent that explores not to chase an external reward but by building a model of its world and acting through it. Its clearest result is biological, showing that a world-model-based intrinsic objective can reproduce an animal’s behavior, and the internal dynamics beneath it, in a real

neural system. Carrying the same idea into visually rich, open-ended environments points to what must still be solved, a world-model substrate that can learn to act, before curiosity can scale to them.

Bibliography

- [1] Yu Mu, Davis V Bennett, Mikail Rubinov, Sujatha Narayan, Chao-Tsung Yang, Masashi Tanimoto, Brett D Mensh, Loren L Looger, and Misha B Ahrens. Glia accumulate evidence that actions are futile and suppress unsuccessful behavior. *Cell*, 178(1):27–43, 2019.
- [2] Reece Keller, Alyn Kirsch Tornell, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Intrinsic goals for autonomous agents: Model-based exploration in virtual zebrafish predicts ethological behavior and whole-brain dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.00138.
- [3] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, 2013.
- [4] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, pages 2778–2787, 2017.
- [5] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.
- [6] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.
- [7] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning (ICML)*, pages 5062–5071, 2019.
- [8] Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International Conference on Machine Learning (ICML)*, pages 5306–5315, 2020.
- [9] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

- [10] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.
- [11] Marion F Haug, Oliver Biehlmaier, Kaspar P Mueller, and Stephan CF Neuhauss. Visual acuity in larval zebrafish: Behavior and histology. *Frontiers in Zoology*, 7:1–7, 2010.
- [12] Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. In *International Conference on Learning Representations (ICLR)*, 2020.
- [13] Diego Aldarondo, Josh Merel, Jesse D Marshall, Leonard Hasenclever, Ugne Klibaite, Amanda Gellis, Yuval Tassa, Greg Wayne, Matthew Botvinick, and Bence P Ölveczky. A virtual rodent predicts the structure of neural activity across behaviors. *Nature*, 2024.
- [14] Roman Vaxenburg, Igor Siwanowicz, Josh Merel, Alice A Robie, Carmen Morrow, Guido Novati, Zinovia Stefanidi, Gert-Jan Both, Gwyneth M Card, Michael B Reiser, et al. Whole-body simulation of realistic fruit fly locomotion with deep reinforcement learning. *bioRxiv*, pages 2024–03, 2024.
- [15] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature Communications*, 14(1):1597, 2023.
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [18] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
- [19] Jenelle Feather, Meenakshi Khosla, N Apurva Ratan Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.
- [20] Luke E. Brezovec, Andrew B. Berger, Shaul Druckmann, and Thomas R. Clandinin. Neural correlates of future volitional action in drosophila. *bioRxiv*, 2023. doi: 10.1101/2023.09.08.556917.

- [21] Victor Lobato-Rios, Shravan Tata Ramalingasetty, Pembe Gizem Özdil, Jonathan Arreguit, Auke Jan Ijspeert, and Pavan Ramdya. Neuromechfly, a neuromechanical model of adult drosophila melanogaster. *Nature Methods*, 19(5):620–627, 2022.
- [22] Sibor Wang-Chen, Victor Alfred Stimpfling, Thomas Ka Chung Lam, Pembe Gizem Özdil, Louise Genoud, Femke Hurtak, and Pavan Ramdya. Neuromechfly v2: Simulating embodied sensorimotor control in adult drosophila. *Nature Methods*, 21:2353–2362, 2024. doi: 10.1038/s41592-024-02497-y.
- [23] György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165–193, 1995.
- [24] Amanda L. Woodward. Infants selectively encode the goal object of an actor’s reach. *Cognition*, 69(1):1–34, 1998.
- [25] Celeste Kidd, Steven T. Piantadosi, and Richard N. Aslin. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, 7(5):e36399, 2012.
- [26] 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)*, pages 8583–8592, 2020.
- [27] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Learning for Dynamics and Control (L4DC)*, pages 761–770. PMLR, 2020.
- [28] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations (ICLR)*, 2024.
- [29] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.
- [30] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*. PMLR, 2023.

Appendix A

Additional Details for Intrinsic Goals in Virtual Zebrafish

This appendix provides additional details for the zebrafish experiments described in Chapter 2. The main text focuses on the conceptual comparison between 3M-Progress and the intrinsic motivation baselines used throughout this thesis. Here, we summarize the computational baselines, the additional control models included in the full comparison, and the behavioral and neural-glial alignment procedures.

Intrinsic Motivation Baseline Details

The main text introduces the intrinsic motivation baselines at the level of intuition. Here, we summarize their computational role and how they differ from 3M-Progress. All baselines generate an intrinsic reward for policy learning in the absence of an external task reward. They differ in which property of the agent’s experience is treated as intrinsically valuable. Prediction-error objectives reward states or transitions that are difficult for a learned forward model to predict. Let ω_θ denote a forward dynamics model that predicts the next sensory feature embedding $\phi(s_{t+1})$ from the current embedding $\phi(s_t)$ and action a_t . A prediction-error reward can be written as

$$r_t^i = \|\phi(s_{t+1}) - \omega_\theta(\phi(s_t), a_t)\|^2. \quad (\text{A.1})$$

The Intrinsic Curiosity Module (ICM) uses this idea in a learned feature space, typically trained with an inverse-dynamics objective so that the representation emphasizes features relevant to the agent’s actions [4]. This objective rewards surprise, but does not distinguish learnable surprise from irreducible noise.

Random Network Distillation (RND) rewards novelty by comparing a trainable predictor network to a fixed randomly initialized target network [6]. Let f be the fixed target network and $\hat{f}_\theta$ the learned predictor. The intrinsic reward is

$$r_t^i = \|\hat{f}_\theta(s_t) - f(s_t)\|^2. \quad (\text{A.2})$$

States that have been visited often produce low prediction error because the predictor has learned the target network’s output there. Rare or unfamiliar states produce higher error and are therefore rewarded. RND is useful as a state-coverage objective, but it does not directly measure whether visiting a state improves the agent’s world model or whether the state is controllable.

Disagreement-based exploration rewards uncertainty across an ensemble of learned forward models [7]. Let $\{\omega_{\theta_k}\}_{k=1}^K$ be an ensemble of predictors. The intrinsic reward is proportional to the variance of their predictions:

$$r_t^i = \text{Var}_k [\omega_{\theta_k}(\phi(s_t), a_t)]. \quad (\text{A.3})$$

When the ensemble agrees, the transition is treated as familiar. When the ensemble disagrees, the transition is treated as informative. This can reduce some failures of raw prediction error, but in complex or stochastic environments the ensemble may continue to disagree even when additional experience is not useful.

Progress-based objectives reward improvement rather than prediction error itself. The general idea is to compare the current model to a previous or lagging version of the model and reward reductions in predictive error. γ -Progress implements this idea with an exponential moving average of the world model, creating a constant-memory approximation to learning progress [8]. In simplified form, if $\bar{\omega}_\theta$ denotes the lagging model and $\omega_{\theta'}$ denotes the current model, the reward measures change in predictive performance across the two:

$$r_t^i \propto |\ell(\bar{\omega}_\theta) - \ell(\omega_{\theta'})|, \quad (\text{A.4})$$

where $\ell(\cdot)$ is a prediction loss on recent experience. This objective avoids rewarding irreducible noise in the same way as raw prediction error, because noisy transitions may remain high-error but do not produce sustained progress. However, because the reference model moves with learning, the reward can vanish once the online model has adapted.

3M-Progress differs from these baselines by comparing the online world model to a fixed ethological memory rather than to a moving copy of itself. The intrinsic reward is based on changes in the mismatch between this fixed memory and the online model. This fixed reference allows the agent to preserve a stable signal in the open-loop environment, where swim commands no longer produce their expected sensory consequences.

Additional Control Models

The main text focuses on comparisons between 3M-Progress and the intrinsic motivation baselines used throughout this thesis: ICM, RND, Disagreement, and γ -Progress. The full model-brain comparison from Keller et al. [2] also includes model-free and data-derived controls. These controls are not the central baselines for the thesis argument, but they help

rule out simpler explanations for the behavioral and neural-glial alignment achieved by 3M-Progress.

The model-free controls test whether the observed behavioral and neural-glial alignment can be explained without a world-model-based intrinsic drive. These agents do not use a learned predictive model of sensory consequences to compute reward. Instead, they test simpler behavioral alternatives.

The *energy-cost*, or homeostatic, control rewards low-effort behavior by penalizing motor output. This control tests whether passivity alone is sufficient to explain the animal-like dynamics. If the zebrafish data were explained simply by reduced movement, then an energy-minimizing agent should align well with the animal. However, this control does not reproduce the structured alternation between active and passive states.

The *maximum-entropy* control encourages broad behavioral variability. This control tests whether high behavioral diversity alone can explain the observed active-passive transitions. Although entropy-based exploration can produce varied behavior, it does not provide a structured signal for when to persist, when to stop, or when to resume activity.

The *randomly initialized* control provides a lower-bound comparison. It tests whether alignment could arise from generic recurrent dynamics or from the architecture alone, without meaningful training. Poor alignment from this control indicates that the match between 3M-Progress and the biological data is not simply a consequence of the network architecture.

The data-derived controls test whether model-brain alignment can be explained by simple dynamical fits or generic statistical structure in the biological recordings, rather than by an intrinsically motivated embodied agent.

The *Gaussian-process* control provides a flexible smooth fit to the observed dynamics. This control tests whether neural-glial alignment can be achieved by fitting the temporal structure of the data without specifying an embodied control objective. Strong performance from this control would suggest that the biological responses are easy to explain with generic smooth dynamics. Instead, 3M-Progress exceeds this control, indicating that the alignment is not explained by temporal smoothing alone.

The *bang-bang PID controller* implements a simplified threshold-based control policy inspired by the proposed biological circuit. This controller switches between activity and passivity using hand-designed control structure rather than learning an intrinsic objective in an embodied environment. It tests whether an explicitly engineered controller can account for the behavioral and neural-glial data. This comparison is useful because it separates fitting the qualitative circuit motif from learning the behavior through self-supervised interaction.

The *whole-brain average response* control uses the average biological activity pattern as a data-derived predictor. This tests whether high alignment can be obtained simply by

matching global activity structure shared across the brain. If the model-brain alignment were driven only by broad population-level dynamics, this control would be expected to perform well. The stronger performance of 3M-Progress indicates that the model captures more specific unit-level dynamics.

The *white-noise random walk* control provides a generic stochastic dynamical baseline. It tests whether random low-frequency fluctuations can align with the slow neural and glial signals observed during active-passive cycling. This control is important because astrocytic signals are slow and could, in principle, be partially matched by unconstrained smooth noise. The comparison shows that 3M-Progress captures structured dynamics beyond random temporal fluctuation.

Together, these controls show that the alignment achieved by 3M-Progress is not explained by passive behavior alone, generic behavioral variability, recurrent architecture, simple temporal smoothing, average whole-brain activity, or unconstrained stochastic dynamics. The model-free controls lack the world-model-based drive needed to reproduce the structured active-passive cycle, while the data-derived controls lack the embodied objective that links behavior to internal dynamics. The central result is therefore not simply that 3M-Progress produces a good fit, but that a fixed ethological memory and an online world model provide a computational structure that explains both the behavior and the neural-glial dynamics observed in larval zebrafish.

Training and Environment Details

Additional implementation details for the zebrafish agent follow Keller et al. [2]. The virtual agent is trained with Proximal Policy Optimization using intrinsic reward only. Across all objectives, a small action penalty is included so that passive behavior remains available to the policy:

$$r_t^a = -\lambda \|a_t\|_2^2. \tag{A.5}$$

This penalty prevents the agent from being forced into continuous movement and makes it possible to evaluate whether each intrinsic objective can produce structured transitions between activity and passivity.

Training proceeds in two stages. First, the ethological memory is learned in the freely swimming environment, where swim commands produce normal sensory consequences. This memory is then frozen. Second, during open-loop training, a separate online world model is learned from the current environment while the policy is optimized using the intrinsic reward produced by the selected objective. For 3M-Progress, the intrinsic reward is computed from the mismatch between the frozen ethological memory and the online world model.

The open-loop environment is designed to reproduce the central manipulation of the zebrafish experiment: the agent can issue swim commands, but those commands no longer produce the expected visual consequences. When the agent is passive, the ethological memory and the online model tend to agree because little self-generated sensory change is expected or produced. When the agent attempts to swim, the models diverge because the ethological memory predicts visual consequences of self-motion while the open-loop environment produces little or no corresponding feedback.

Behavioral Alignment Details

Behavioral alignment is computed by comparing event-triggered swim-power traces around active and passive transitions. In zebrafish, swim power is derived from motor nerve recordings. In the virtual agent, swim power is derived from the norm of joint torques. These quantities are not physically identical, but both provide a time series of motor output. For this reason, alignment is evaluated on the temporal dynamics around transitions rather than on absolute motor scale.

The main behavioral criterion is whether an agent reproduces the active-passive cycling observed in the animal. This requires more than matching an average level of movement. A model can have the same mean activity as the animal while failing to reproduce the long-timescale alternation between active swimming and passive pauses. The alignment analysis therefore focuses on the temporal structure of behavior around transitions and compares model behavior to the inter-animal ceiling defined by alignment between zebrafish individuals.

Neural-Glial Alignment Details

Model-brain alignment is computed by aligning biological calcium activity and artificial unit activity to the same behavioral transition windows. The biological data include whole-brain recordings of neural and glial activity during futility-induced passivity [1]. Model activations are taken from the trained agent during the corresponding open-loop behavior. Both biological and model activities are aligned to active and passive transitions using event-triggered averages.

Each biological unit is matched to the single most correlated artificial unit. This one-to-one mapping is intentionally restrictive. Unlike linear regression or other many-to-one mappings, it does not allow arbitrary combinations of artificial units to explain a biological response. Instead, it asks whether individual artificial units exhibit dynamics comparable to individual neural or glial units. Correlations are noise-corrected using the inter-animal reliability ceiling, which reflects the degree to which biological units align across zebrafish individuals.

This alignment procedure tests whether the model captures the internal computation underlying futility-induced passivity, not only the outward behavior. A model could reproduce

active–passive cycling through a control strategy unrelated to the biological circuit, or it could produce internal dynamics that resemble neural activity without generating the correct behavior. The NeuroAI Turing Test criterion applied in the main text requires both behavioral and internal alignment [19]. 3M-Progress is the only tested model that approaches the inter-animal ceiling on both axes in the zebrafish dataset.

Functional Mapping of 3M-Progress to the Neural-Glial Circuit

The alignment analysis is motivated by a functional correspondence between components of 3M-Progress and the neural-glial circuit identified in larval zebrafish. The instantaneous model-memory mismatch ϵ_t has the computational role of a failed-action signal. It increases when the agent attempts to swim but the expected sensory consequence does not occur, making it functionally analogous to NE-MO neurons that respond to failed swim attempts. The filtered mismatch $\hat{\epsilon}_t$ has the computational role of an accumulated failure signal. Because it integrates recent mismatch through a leaky update, it is functionally analogous to radial astrocytes that accumulate noradrenergic input over repeated failed swim attempts and decay during passivity. The reward derived from deviations between ϵ_t and $\hat{\epsilon}_t$ then encourages transitions between active behavior, where mismatch is revealed, and passive behavior, where mismatch is reduced.

This correspondence is functional rather than mechanistic. The model does not claim that artificial units are biologically identical to NE-MO neurons or radial astrocytes. Instead, it tests whether an intrinsic objective with the same computational structure can reproduce both the behavioral transition and the neural-glial dynamics observed in the animal.