Dissecting Zebrafish Hunting Behavior using Deep Reinforcement Learning trained RNNs

Raaghav Malik *
California Institute of Technology
rmalik@caltech.edu

Satpreet H. Singh *
Harvard University

Sonja Johnson-Yu *
Harvard University

Roy Harpaz
Harvard University

Kanaka Rajan
Harvard University

Abstract

Larval zebrafish hunting provides a tractable setting to study how ecological and energetic constraints shape adaptive behavior in both biological brains and artificial agents. Here we develop a minimal agent-based model, training recurrent policies with deep reinforcement learning in a bout-based zebrafish simulator. Despite its simplicity, the model reproduces hallmark hunting behaviors—including eye vergence-linked pursuit, speed modulation, and stereotyped approach trajectories—that closely match real larval zebrafish. Quantitative trajectory analyses show that pursuit bouts systematically reduce prey angle by roughly half before strike, consistent with measurements. Virtual experiments and parameter sweeps vary ecological and energetic constraints, bout kinematics (coupled vs. uncoupled turns and forward motion), and environmental factors such as food density, food speed, and vergence limits. These manipulations reveal how constraints and environments shape pursuit dynamics, strike success, and abort rates, yielding falsifiable predictions for neuroscience experiments. These sweeps identify a compact set of constraints—binocular sensing, the coupling of forward speed and turning in bout kinematics, and modest energetic costs on locomotion and vergence—that are sufficient for zebrafish-like hunting to emerge. Strikingly, these behaviors arise in minimal agents without detailed biomechanics, fluid dynamics, circuit realism, or imitation learning from real zebrafish data. Taken together, this work provides a normative account of zebrafish hunting as the optimal balance between energetic cost and sensory benefit, highlighting the trade-offs that structure vergence and trajectory dynamics. We establish a *virtual lab* that narrows the experimental search space and generates falsifiable predictions about behavior and neural coding.

1 Introduction

Adaptive behavior unfolds under ecological and energetic constraints that shape what strategies are feasible or optimal [1, 2]. Understanding how such constraints give rise to structured behavioral sequences is a shared challenge for neuroscience, neuroAI, and artificial intelligence [3, 4].

Larval zebrafish hunting is a particularly clear example of structured behavior: animals pursue prey through discrete bouts organized into exploration, orientation, pursuit, and either strike or abort [5, 6, 2]. These behaviors exhibit consistent hallmarks, including a vergence-linked shift into a "hunting mode" [5, 6], systematic halving of prey angle across pursuit bouts [2], and stereotyped

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Reach and Limits of AI for Scientific Discovery.

approach trajectories [6]. Despite this well-documented structure, it remains unclear why these behavioral motifs emerge and persist, or under what constraints they represent optimal solutions [1]. In particular, we lack explanations for why vergence angles shift abruptly [6], why prey angle is consistently reduced by about 50% per pursuit bout [2], and why some hunts succeed while others abort [1]. Prior computational accounts, including bounded integrator models [7] and probabilistic inference frameworks [2], capture aspects of zebrafish hunting but stop short of explaining why stereotyped trajectories emerge as optimal strategies.

Models of behavior in neuroscience are often *descriptive*, specifying what actions animals perform under given conditions [6, e.g.,], or *mechanistic*, capturing how neural circuits generate those behaviors [8, e.g.,]. In contrast, our approach is *normative*: it explains *why* a behavioral strategy emerges as optimal given specified constraints, in line with reinforcement-learning–based normative models of perception and decision-making [7, 9, e.g.,].

Although larval zebrafish are unusually accessible for circuit-level and behavioral experiments thanks to their transparency and genetic tools [10, 11], careful ethological work still leaves key variables hard to isolate and manipulate [6, 1]. Virtual-reality assays can reliably evoke components, such as convergent eye movements and orienting turns, but typically do not permit fine-grained, closed-loop control over the entire hunting sequence [5]. In naturalistic arenas, ecological variables such as prey density, prey kinematics, and energetic costs covary, making it difficult to vary them systematically one at a time for causal inference. Moreover, internal state variables—such as motivational drive or accumulated evidence—that likely govern the transition between pursuit, strike, and abort remain difficult to access with current experimental methods [7, 12].

Task-optimized artificial neural networks provide a complementary approach, offering a way to test how specific constraints give rise to structured behavior when direct experimentation falls short [9, 3]. Prior studies show that recurrent neural network (RNN) agents trained with deep reinforcement learning (DRL) can capture biological strategies, such as electrosensory navigation in weakly electric fish [13, 14]. The same methods have also been applied in AI and NeuroAI settings, where they give rise to complex planning behaviors and structured internal representations [4, 15, 16, 17, 18, 19, 20]. Taken together, this body of work motivates applying task-optimized recurrent agents to zebrafish hunting as a principled way to probe how ecological and energetic constraints shape structured behavior.

Here, we introduce a biologically inspired hunting simulator where RNN-based DRL agents learn to pursue prey through discrete bouts (with prey modeled as stochastic walkers mimicking paramecia rather than adversarial agents [21]). This framework enables systematic manipulations of ecological variables, sensory constraints, and energetic costs, providing a virtual laboratory for uncovering how constraints yield structured behavior. The same approach—task-optimized DRL agents analyzed via structured sweeps—can extend beyond zebrafish hunting to other sensorimotor systems and inform inductive biases in AI and robotics.

Our key contributions are:
(i) We introduce a biologically inspired *framework* in which virtual zebrafish agents perceive, move, and hunt through discrete bouts, enabling systematic manipulation of ecological, sensory, and energetic constraints.
(ii) We train recurrent agents with deep reinforcement learning and show that they spontaneously develop naturalistic hunting strategies without imitation learning from real zebrafish data.
(iii) Through detailed behavioral analyses and parameter sweeps, we identify a compact set of constraints—binocular sensing, bout kinematics, and energetic costs—minimal yet sufficient for zebrafish-like hunting behavior to emerge.
(iv) We establish a *virtual lab* that provides falsifiable predictions for in vivo experiments, offers a normative account of why hunting stereotypy emerges, and serves as a starting point for probing how task-trained RNNs might encode behavioral state variables.

2 Methods

2.1 Simulation Environment and Agent Design

Following experimental studies of larval zebrafish in open circular arenas [7, 22], we simulate a two-dimensional circular aquatic arena with rigid boundaries and diameter between 33 and 100 mm.

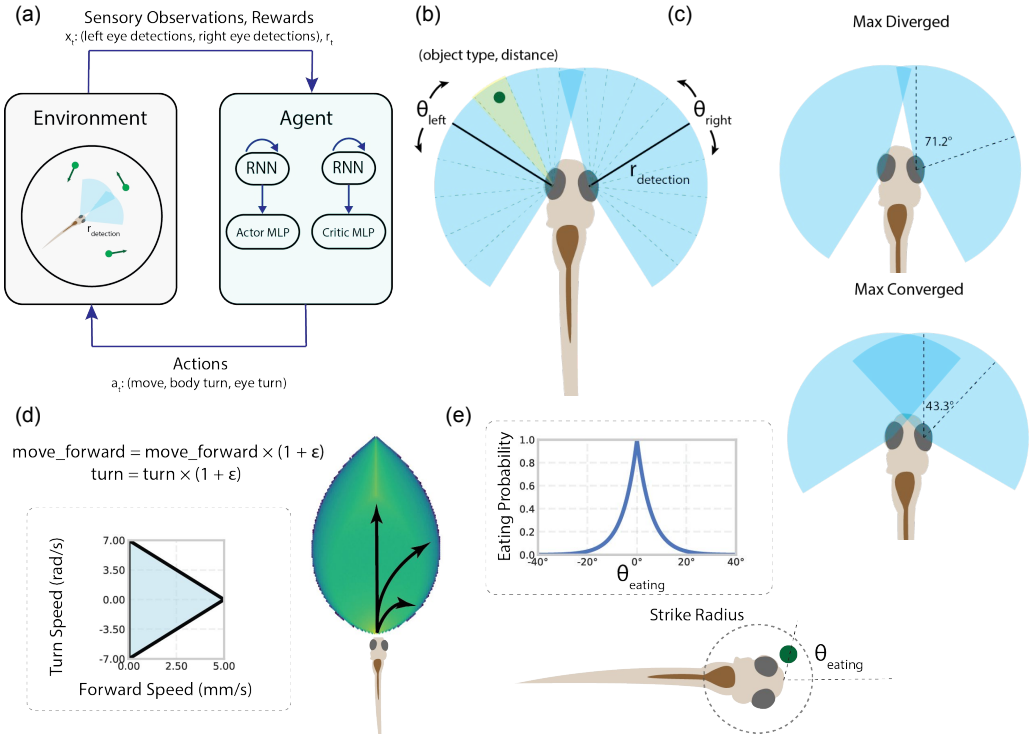

Figure 1: (a) Closed-loop RL framework: an actor–critic RNN agent selects actions that update the environment state, which in turn provides new observations and rewards. (b) Zebrafish agent model. Each eye has a 163 degree perception field that can rotate, and is divided into 10 sectors reporting the type and distance of the nearest object. (c) Eye configurations at maximum divergence and convergence; the agent can adopt any intermediate vergence angle. (d) The agent's action space, with coupled forward and turn speeds. Independent multiplicative uniform noise is added to both. (e) Prey capture: when prey are within strike distance, capture probability depends on angular alignment to the food, modeled by a Laplace distribution.

The arena contains a fixed number of stochastically moving prey agents, with both the prey and the zebrafish agent initialized uniformly at random within the arena (Figure 1a, left). Prey motion statistics are chosen to approximate *Paramecium* swimming behavior observed in feeding experiments [6]. The agent (Fig 1a, right) consists of two recurrent neural networks (RNN) of 256 units each [8] and parallel two-layer Actor and Critic Multi-Layer Perceptrons (MLPs) [23]. The agent acts in closed-loop with the environment, generating actions that change the environment state, which in turn provides sensory observations and rewards.

We model the agent with two eyes with a fixed forward offset and width, with each eye having a monocular field of view of 163 degrees and a detection radius of 10 mm matched to empirical larval zebrafish visual field experiments [5]. Each eye's field of view is divided into 10 angular sectors. The sectors widen with radial distance, modeling reduced spatial resolution (Figure 1b. Each sector receives object type and distance information to the closest object (food, wall) in that sector. The agent can rotate its eyes to change the vergence angle, which defines the size of the binocular region. The maximum and minimum vergence values are set according to experimental data (Figure 1c).

We model each action as a single bout, following a similar approach to [6], with each bout taking 125 ms. The forward speed and turn speed of each bout are coupled. Larger forward speeds permit only smaller turns, creating a triangular action space (Figure 1d). Multiplicative noise is then added independently to the forward and turn speeds. When the agent is within strike distance of food, the probability of the strike being successful is related to the angle to the food θ_{eating} via a Laplace distribution (Figure 1e), with decay rate chosen to match the empirical distribution of prey angles when larval zebrafish perform strike bouts [2].

2.2 Configurable Agent Features

We develop a framework where various agent features can be toggled and tuned to determine what features result in the emergence of naturalistic zebrafish hunting behavior as an optimal strategy. This, in turn, provides a normative explanation for why larval zebrafish exhibit a distinctly stereotyped hunting mode. The key features are:

- **Perception noise:** We introduce multiplicative perception noise in the distance estimates of the eye sensors. Each sensor has a fixed probability of producing false positives (detecting an object when none exists) and false negatives (missing an existing object). This makes the binocular region advantageous, since two independent observations are collected for each object instead of one in the monocular region. In an extreme case, we restrict distance information to the binocular region, with monocular sensing limited to detection only. We also allow angular noise to be tuned in the monocular region; in the most extreme setting, the firing sector of the eye is chosen uniformly at random.
- **Speed cost:** A ReLU-like energy penalty for large forward speeds above a fixed threshold, reflecting the fact that sustained high-speed swimming is energetically costly and physiologically limited in larval zebrafish.
- **Vergence cost:** An energy penalty for eye vergence deviation from rest (divergence), reflecting the fact that sustaining the convergence state requires flexing the eye muscle and is costly for larval zebrafish.
- **Eye fatigue:** We model eye fatigue as $f_{eye} = |\Delta\theta_{right}| + |\Delta\theta_{left}|$ where $\Delta\theta_{left}$ and $\Delta\theta_{right}$ are the per-bout changes in left and right eye angles. A linear penalty is applied once f_{eye} exceeds a threshold equal to one full sweep between divergence and convergence.
- **Eye control:** Agents can be configured with either independently controlled eyes or coupled eyes with mirrored vergence angles.

For the below results, we only provide distance and angle information in the binocular region, use one degree of freedom eyes, and do not use eye fatigue. All other constants used are given in Appendix 5.1.

2.3 Reward Structure and Training

The primary reward is for successful prey capture, with much smaller penalties for energetic costs and fatigue as mentioned above. An even smaller distance-based shaping reward is also used during training to encourage prey approach for initial convergence.

The per-timestep reward is given by:

$$\begin{aligned} R_t = R_{\text{capture}} &- \alpha_{\text{eye}} \left(\|\theta_{\text{left}} - \theta_{\text{left, rest}}\| + \|\theta_{\text{right}} - \theta_{\text{right, rest}}\| \right) \\ &- \beta_{\text{speed}} \cdot \max(v_{\text{current}} - v_{\text{threshold}}, 0) \\ &- \gamma_{\text{fatigue}} \max(f_{\text{eye}} - f_{\text{threshold}}, 0) + R_{\text{shape}} \end{aligned} \tag{1}$$

where α_{eye}, β_{speed}, and γ_{fatigue} are constants chosen such that energy penalties are 1 order of magnitude smaller than food reward and shaping rewards are 2 orders of magnitude smaller. For all parameter values and chosen features, see Appendix 5.1.

We train an *actor-critic recurrent neural network* policy using Proximal Policy Gradient (PPO) [24] in a single-agent setting. Training follows a curriculum in which prey density is gradually reduced, prey motion becomes more variable, and the strike probability distribution is tightened, requiring increasingly precise strikes.

3 Results

3.1 Hunting and Exploration Show Distinct Trajectories

A hunting sequence begins when a prey item is first detected and ends either with a capture (strike) or when the prey exits the perception radius without capture (abort). All other periods are defined as exploration.

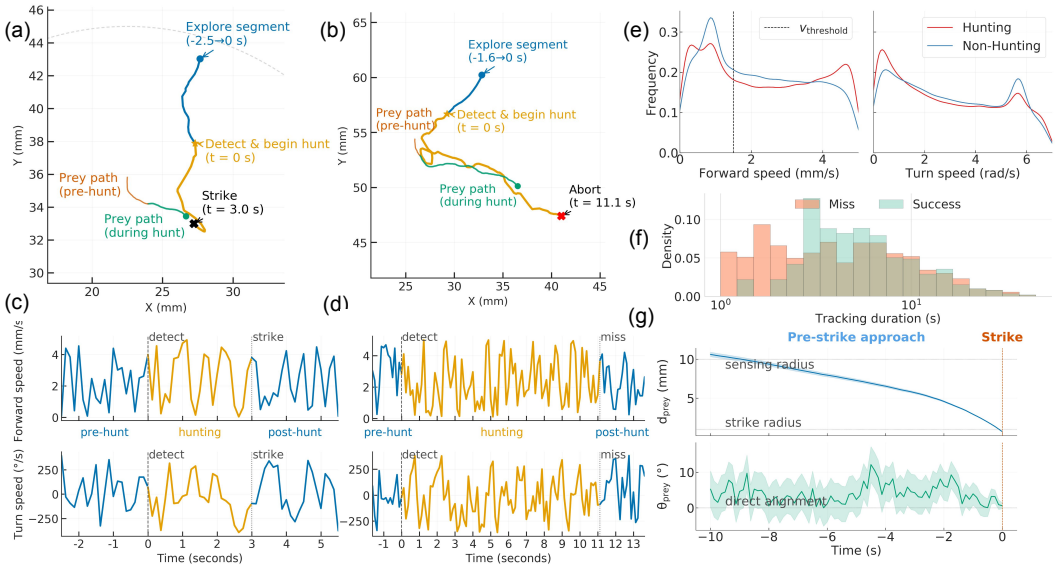

Figure 2: (a) Example trajectory that ends in a successful strike. (b) Example trajectory that ends in an abort. (c–d) Forward speed and orientation over time for the trajectories in (a) and (b). (e) Distribution of forward and turn speeds during hunting versus exploration. (f) Hunt duration, separated by successes and failures. (g) Average distance and angle to prey preceding successful strikes.

Figure 2a and 2b show sample hunting trajectories that end in a successful strike and an abort, respectively, with Figure 2c and 2d showing the forward speed and turn speed before, during, and after these sample hunts. We find that the hunt trajectories mirror those empirically found in larval zebrafish with alternating periods of fast and slow motion, much like zebrafish bouts [6].

3.2 Movement Statistics Differentiate Successful and Failed Hunts

Figure 2e depicts the agent's forward speed and turn speed distributions when hunting vs. not hunting, evaluated across a set of 200 fixed arena sizes, food locations, and agent initializations. During hunting, the agent more frequently exceeds the forward speed threshold, whereas it tends to remain below threshold during exploration. This suggests that the agent accepts higher movement costs in pursuit of food reward, consistent with empirical findings [6]. In addition, the agent executes smaller and more precise turns while hunting than while exploring.

We see that successful strikes have a longer tracking duration on average than unsuccessful strikes (Figure 2f). This suggests that the agent implicitly represents the likelihood of success and aborts early when that probability is low, choosing instead to explore for other targets.

Finally, across all successful hunting sequences, the agent's average distance to the prey and its angular alignment error both decrease consistently (Figure 2g).

3.3 Eye Convergence During Hunting Mirrors Empirical Zebrafish Data

Agents show higher eye vergence during successful hunts compared to non-hunting periods (Figure 3a). During hunting, vergence angle increases steadily and peaks just before the strike (Figure 3b), closely matching empirical zebrafish data [5]. This suggests that the agent accepts the energetic cost of deviating from the resting eye position in exchange for improved sensory information.

Before and immediately after prey detection, vergence angles are similar in successful and failed hunts. Midway through tracking, however, the trajectories diverge: successful hunts continue to increase vergence, while failed hunts do not. This supports the view that the agent represents the likelihood of success and invests in the costly vergence increase only when the hunt is likely to succeed.

Figure 3c shows vergence trajectories for five successful hunts. Figure 3d summarizes the proportion of time food is located in binocular versus monocular regions across all successful hunts.

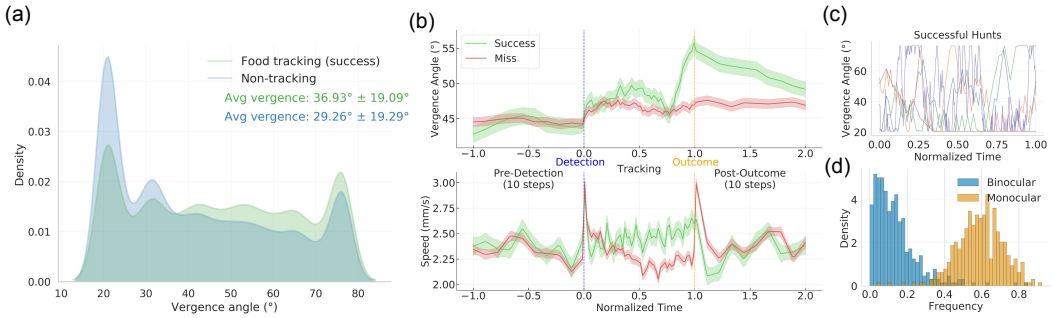

Figure 3: (a) Distribution of vergence angles during successful hunts versus non-hunting periods. (b) Time course of vergence angle before prey detection, during tracking, and after strike or abort, for both successful and failed hunts. (c) Example vergence trajectories for individual successful and failed hunts. (d) Proportion of time food is located in binocular versus monocular regions across all successful hunts.

3.4 RNN Hidden State encodes behavior-relevant variables

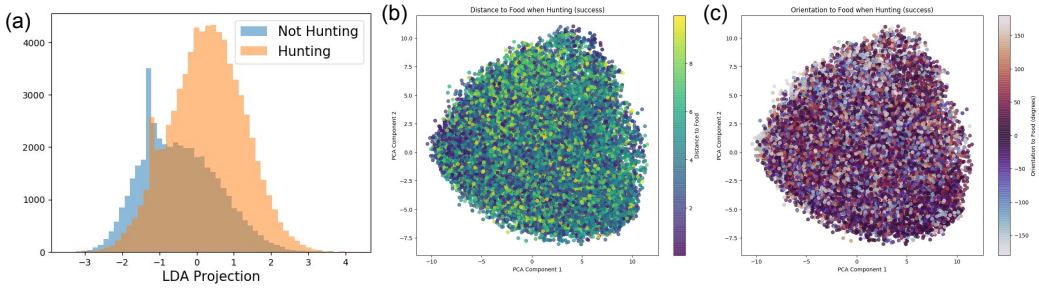

Figure 4: (a) LDA projection of RNN hidden states, showing separation between hunting and non-hunting periods. (b-c) RNN Hidden states projected onto the first two principal components (PCs) during successful hunts, colored by (b) prey distance and (c) prey orientation.

An LDA projection of the RNN hidden states separates hunting from non-hunting periods, indicating that the network encodes the hunting state (Figure 4a). A logistic decoder trained on these projections achieves a classification accuracy of 0.61. We also check for distance and orientation encoding (Figures 4b and 4c).

4 Discussion

We show that a recurrent neural network agent, trained with deep reinforcement learning in a biologically grounded zebrafish hunting environment, spontaneously develops multiple hallmark features of larval zebrafish hunting behavior. These include (i) a discrete shift into a high-vergence "hunting mode" following prey detection, (ii) stereotyped approach trajectories with smooth decreases in prey angle and distance, and (iii) speed modulation and precise turning during pursuit. Notably, these behaviors emerge from optimizing a reward function with only sparse capture rewards and modest energy penalties, without hard-coding the structure of the hunting sequence.

Our model offers a normative explanation for discrete vergence shifts. Despite their energy cost, convergence is favored in high-confidence hunts because binocular sensing improves prey localization. Coupled turn–move constraints, multiplicative action noise, and uncertainty in prey movement further incentivize convergence. Our experiments suggest that the clear vergence switch in the hunting state is driven not by the increased perceptual angle of the divergence state, but by the tradeoff between the

energetic cost of maintaining convergence and the perceptual benefits it provides. It is also possible that unmodeled factors, such as drives to track conspecifics or avoid predators, contribute to the behavioral differences between our in-silico agents and real zebrafish larvae. Similar trade-offs may underlie the evolution and robustness of hunting stereotypy in larval zebrafish.

Higher forward speeds during hunting also reflect a cost–benefit balance between capture probability and energy expenditure. The observation that low-probability hunts are aborted early suggests an implicit value-based decision process, paralleling foraging strategies across species.

Hidden-state analyses reveal a clear separation between hunting and exploration, as well as continuous representations of prey distance and orientation. This supports the existence of an internal "hunting mode" variable. Further work is needed to disentangle these encodings from raw observations and to test whether the network learns a predictive model of prey motion.

Our agent's vergence trajectories, pursuit kinematics, and bout structure are strikingly consistent with experimental reports [5, 6, 2]. This convergence, despite substantial differences in physical embodiment and neural implementation,supports the idea that these strategies are optimal solutions given the zebrafish's sensory constraints and ecological demands. The agreement also reinforces the utility of task-optimized neural networks as normative models for animal behavior [3].

At the same time, our model omits important biological constraints, including multisensory integration, detailed biomechanics, and anatomically grounded connectivity. Incorporating these features could enable more precise circuit-level predictions.

More broadly, this work illustrates how deep RL with recurrent policies in ethologically inspired simulations can bridge normative and mechanistic accounts of behavior. By tuning environmental parameters and agent costs, we can ask "why" a behavior is structured as it is, and "how" it could be implemented in a recurrent network with biological constraints. This approach also complements experimental work by allowing controlled manipulations of variables that are difficult or impossible to isolate in vivo, generating testable predictions for neural coding, internal state dynamics, and behavioral strategies.

Acknowledgements

We thank members of the Rajan and Engert labs for helpful discussions. Funded by NIH (RF1DA056403 to K.R., U19NS104653), James S. McDonnell Foundation (220020466), Simons Foundation (Pilot Extension-00003332-02 to K.R.), McKnight Endowment Fund (K.R.), CIFAR Azrieli Global Scholar Program (K.R.), NSF (2046583 to K.R.), Harvard Medical School Neurobiology Lefler Small Grant Award (K.R.), Harvard Medical School Dean's Innovation Award (K.R. and S.H.S.), Caltech Summer Undergraduate Research Fellowship (R.M.), and Alice and Joseph Brooks Fund Postdoctoral Fellowship (S.H.S.).

References

[1] Shuyu I Zhu and Geoffrey J Goodhill. From perception to behavior: The neural circuits underlying prey hunting in larval zebrafish. *Front. Neural Circuits*, 17:1087993, February 2023.

[2] Andrew D Bolton, Martin Haesemeyer, Josua Jordi, Ulrich Schaechtle, Feras A Saad, Vikash K Mansinghka, Joshua B Tenenbaum, and Florian Engert. Elements of a stochastic 3d prediction engine in larval zebrafish prey capture. *eLife*, 8:e51975, nov 2019. ISSN 2050-084X. doi: 10.7554/eLife.51975. URL `https://doi.org/10.7554/eLife.51975`.

[3] Satpreet Harcharan Singh. Neuroprospecting with DeepRL agents. NeurIPS 2021 Workshop on AI for Science, 2021.

[4] Ann Huang, Satpreet Harcharan Singh, and Kanaka Rajan. Learning dynamics and the geometry of neural dynamics in recurrent neural controllers. In *Workshop on Interpretable Policies in Reinforcement Learning@ RLC-2024*.

[5] Isaac H. Bianco, Adam R. Kampff, and Florian Engert. Prey capture behavior evoked by simple visual stimuli in larval zebrafish. *Frontiers in Systems Neuroscience*, volume 5 - 2011, 2011. ISSN 1662-5137. doi: 10.3389/fnsys.2011.00101. URL `https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2011.00101`.

[6] Robert Evan Johnson, Scott Linderman, Thomas Panier, Caroline Lei Wee, Erin Song, Kristian Joseph Herrera, Andrew Miller, and Florian Engert. Probabilistic models of larval zebrafish behavior reveal structure on many scales. *Curr. Biol.*, 30(1):70–82.e4, January 2020.

[7] Armin Bahl and Florian Engert. Neural circuits for evidence accumulation and decision making in larval zebrafish. *Nat. Neurosci.*, 23(1):94–102, January 2020.

[8] Kanaka Rajan, Christopher D Harvey, and David W Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128–142, April 2016.

[9] Martin Haesemeyer, Alexander F Schier, and Florian Engert. Convergent temperature representations in artificial and biological neural networks. *Neuron*, 103(6):1123–1134.e6, September 2019.

[10] Owen Randlett, Caroline L Wee, Eva A Naumann, Onyeka Nnaemeka, David Schoppik, James E Fitzgerald, Ruben Portugues, Alix M B Lacoste, Clemens Riegler, Florian Engert, and Alexander F Schier. Whole-brain activity mapping onto a zebrafish brain atlas. *Nat. Methods*, 12(11):1039–1046, November 2015.

[11] Claire Leyden, Christian Brysch, and Aristides B Arrenberg. A distributed saccade-associated network encodes high velocity conjugate and monocular eye movements in the zebrafish hindbrain. *Sci. Rep.*, 11 (1):12644, June 2021.

[12] Owen Randlett, Martin Haesemeyer, Greg Forkin, Hannah Shoenhard, Alexander F Schier, Florian Engert, and Michael Granato. Distributed plasticity drives visual habituation learning in larval zebrafish. *Curr. Biol.*, 29(8):1337–1345.e4, April 2019.

[13] Sonja Johnson-Yu, Satpreet Harcharan Singh, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Understanding biological active sensing behaviors by interpreting learned artificial agent policies. In *Workshop on Interpretable Policies in Reinforcement Learning@ RLC-2024*, 2024.

[14] Satpreet H Singh, Sonja Johnson-Yu, Zhouyang Lu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel B Sawtell, and Kanaka Rajan. Understanding electro-communication and electro-sensing in weakly electric fish using multi-agent deep reinforcement learning. *arXiv preprint arXiv:2511.08436*, 2025.

[15] Ann Huang, Satpreet H Singh, Flavio Martinelli, and Kanaka Rajan. Measuring and controlling solution degeneracy across task-trained recurrent neural networks. *ArXiv*, pages arXiv–2410, 2025.

[16] Riley Simmons-Edler, Ryan P Badman, Felix Baastad Berg, Raymond Chua, John J Vastola, Joshua Lunger, William Qian, and Kanaka Rajan. Deep RL needs deep behavior analysis: Exploring implicit planning by model-free agents in open-ended environments. 2025.

[17] Satpreet H Singh, Floris van Breugel, Rajesh PN Rao, and Bingni W Brunton. Emergent behaviour and neural dynamics in artificial agents tracking odour plumes. *Nature machine intelligence*, 5(1):58–70, 2023.

[18] Reece Keller, Alyn Tornell, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Autonomous behavior and whole-brain dynamics emerge in embodied zebrafish agents with model-based intrinsic motivation, 2025. URL `https://arxiv.org/abs/2506.00138`.

[19] Satpreet Harcharan Singh, Sonja Johnson-Yu, Zhouyang Lu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Proposal: Deciphering electrocommunication with marl and unsupervised machine translation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*, 2025.

[20] Kaden Zheng, Sonja Johnson-Yu, Satpreet Harcharan Singh, Denis Turcu, Federico Pedraja, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Keypoint annotation for electrocommunication source separation with pikachu and raichu. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*, 2025.

[21] Dhruv Zocchi, Millen Nguyen, Emmanuel Marquez-Legorreta, Igor Siwanowicz, Chanpreet Singh, David A. Prober, Elizabeth M. C. Hillman, and Misha B. Ahrens. Days-old zebrafish rapidly learn to recognize threatening agents through noradrenergic and forebrain circuits. *Current Biology*, 35(1):163–, 2025. doi: 10.1016/j.cub.2024.11.057. PMID: 39719697.

[22] Roy Harpaz, Ariel C Aspiras, Sydney Chambule, Sierra Tseng, Marie-Abèle Bind, Florian Engert, Mark C Fishman, and Armin Bahl. Collective behavior emerges from genetically controlled simple behavioral motifs in zebrafish. *Sci. Adv.*, 7(41):eabi7460, October 2021.

[23] Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. 2021.

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

[25] Ansa W Fiaz, Karen M Léon-Kloosterziel, Gerrit Gort, Stefan Schulte-Merker, Johan L van Leeuwen, and Sander Kranenbarg. Swim-training changes the spatio-temporal dynamics of skeletogenesis in zebrafish larvae (danio rerio). *PLoS One*, 7(4):e34072, April 2012.

[26] Cees J Voesenek, Remco P M Pieters, Florian T Muijres, and Johan L van Leeuwen. Reorientation and propulsion in fast-starting zebrafish larvae: an inverse dynamics analysis. *J. Exp. Biol.*, 222(Pt 14): jeb203091, July 2019.

[27] Biswadeep Khan, On-Mongkol Jaesiri, Ivan P Lazarte, Yang Li, Guangnan Tian, Peixiong Zhao, Yicheng Zhao, Viet Duc Ho, and Julie L Semmelhack. Zebrafish larvae use stimulus intensity and contrast to estimate distance to prey. *Curr. Biol.*, 33(15):3179–3191.e4, August 2023.

5 Appendix

5.1 Numerical values used

Table 1: Simulation parameters.

(a) Fish constants used in the simulation. Some values are approximated to simplify empirical distributions.

Parameter	Value	Unit	Description / Notes
max_speed	5	mm/s	Max speed of larval zebrafish when foraging (approx from [25])
max_turn_speed	7	rad/s	Max turning speed of larval zebrafish when foraging (approx from [26])
max_eye_turn_speed	0.8	rad/s	Max eye turning speed (approx from [11])
perception_field	$163 \cdot \pi/180$	rad	Monocular field of view [5]
max_left_vergence	$-43.3 \cdot \pi/180$	rad	Left eye at maximum convergence [5]
max_right_vergence	$43.3 \cdot \pi/180$	rad	Right eye at maximum convergence [5]
min_left_vergence	$-71.2 \cdot \pi/180$	rad	Left eye at maximum divergence [5]
min_right_vergence	$71.2 \cdot \pi/180$	rad	Right eye at maximum divergence [5]
bout_length	0.125	s	Duration of a bout [2, 6]
eye_separation	2	mm	Distance between eyes (approx from [5])
eye_forward_offset	0.5	mm	Forward offset of the eyes from the center of the agent (approx from [5])
detection_range	10	mm	Max food/wall detection range (at noisy, lowest resolution) (approx from [1])
eating_distribution_decay	10	–	Laplace decay for strike probability vs. orientation (fit to [6])
eating_angle	$80 \cdot \pi/180$	rad	Cutoff half-angle for strike probability (fit to [6])
strike_radius	1	mm	Strike distance [27]
distance_noise_std	0.01	–	Std. of uniform multiplicative noise per sensor
detection_failure_rate	0.0	–	False negative rate per sensor
false_positive_rate	0.0	–	False positive rate per sensor
penalize_move_threshold	1.5	mm/s	Threshold for ReLU-like penalty on forward speed
action_noise_std	0.0	–	Uniform multiplicative noise on forward/turn speeds

(b) Prey-related environment parameters.

Parameter	Value	Unit	Description / Notes
food_speed	1	mm/s	Speed of paramecia
food_turn_std	$10 \cdot \pi/180$	rad/s	Std. of (uniform) turn per step of paramecia
food_density	0.003	count/mm^2	Density of paramecia in arena

(c) Reward parameters.

Parameter	Value
R_{capture}	10
α_{eye}	0.005
β_{speed}	0.01
γ_{fatigue}	0
R_{shape}	~ 0.01