
Visual Robustness and Neural Alignment in a Shared Foraging Task: The Mouse vs. AI Benchmark

Marius Schneider^{1,*,&}, Joe Canzano^{2,*}, Yuchen Hou^{3,*}, Jing Peng^{3,*}, Anjali Deepu⁴, Utsab Karan⁴, Phu-Hoa Pham⁵, Chi-Nguyen Tran⁵, Dao Sy Duy Minh⁵, Nguyen Lam Phu Quy⁵, Huynh Trung Kiet⁵, Simone Azeglio⁶,
Spencer LaVere Smith^{2,&,&}, Michael Beyeler^{3,4,&}

¹ Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA, USA, ² Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA, ³ Department of Computer Science, University of California, Santa Barbara, CA, USA, ⁴ University of Science, VNU-HCM, Ho Chi Minh City, Vietnam, ⁵ Indian Institute of Technology, Kharagpur, India ⁶ Institut de la Vision, Sorbonne Universite, Paris, France, ⁷ Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA. * co-first authors. & co-senior authors. & Correspondence: marius_schneider@ucsb.edu; sls@ucsb.edu

Abstract

1 Visual robustness under real-world conditions remains a critical bottleneck for
2 modern reinforcement learning agents. In contrast, biological systems such as
3 mice show remarkable resilience to environmental changes, maintaining stable
4 performance even under degraded visual input with minimal exposure. Inspired by
5 this gap, we established the Mouse vs. AI: Robust Foraging Competition, a novel
6 bioinspired visual robustness and neural alignment benchmark in which agents and
7 mice perform the same naturalistic 3D foraging task. The benchmark consists of
8 two tracks: (1) a robustness track evaluating generalization to unseen visual pertur-
9 bations, and (2) a neural alignment track measuring how well model representations
10 predict large-scale neural recordings from mouse visual cortex. We provide a public
11 benchmark suite comprising a Unity-based environment for training and evaluating
12 virtual agents, centralized evaluation protocols, and an extensive neural dataset
13 with approximately 178 minutes of multimodal recordings of behavior, visual
14 input, and neural activity from more than 50,000 neurons in mice performing the
15 task. As part of the NeurIPS competition, we received submissions from 22 teams
16 proposing diverse architectures and training strategies. By embedding evaluation in
17 a shared sensorimotor task performed by both biological and artificial agents, the
18 benchmark enables a class of evaluative claims not supported by existing robustness
19 or neural alignment benchmarks in isolation: namely, whether robust, brain-like
20 visual representations emerge spontaneously from behavior-driven learning, and
21 whether robustness and neural alignment reinforce or trade off against one another.
22 Access to the dataset as well as the benchmarking infrastructure will remain online
23 at <https://robustforaging.github.io/>.

24 1 Introduction

25 Generalization under distributional shift remains a fundamental challenge in machine learning. Visual
26 systems trained under standard conditions often fail when exposed to perturbations such as fog,
27 dynamic lighting, or environmental clutter [Wang et al., 2024]. This brittleness limits the deployment
28 of vision-based autonomous systems in safety-critical scenarios, including self-driving vehicles [Lu
29 et al., 2023], drones [Gao et al., 2021], and service robots operating in real-world environments.

In contrast, biological systems exhibit remarkable visual robustness within their ecological and sensorimotor niches. Mice, for example, can navigate complex environments and perform visually guided tasks even when visual input is degraded. This discrepancy is unlikely to be explained solely by superior sensors, as biological eyes lag behind modern cameras in resolution and precision, but instead suggests that the visual system implements computational principles that support invariant and noise-tolerant processing.

Recent work has developed complementary benchmarks to study artificial and biological vision. Brain-Score [Schrimpf et al., 2018] evaluates the alignment between model representations and neural activity across species and brain regions, while Sensorium [Willeke et al., 2022, Turishcheva et al., 2024] focuses on predicting neural responses from large-scale recordings in mouse visual cortex. In parallel, robustness benchmarks in computer vision, such as ImageNet-C [Hendrycks and Dietterich, 2019] and the Adversarial Vision Challenge [Brendel et al., 2020], evaluate generalization under distributional shift. However, these approaches study robustness and neural alignment largely in isolation.

To bridge this gap, we introduce the **Mouse vs. AI: Robust Foraging** benchmark and competition, a biologically grounded evaluation suite that jointly evaluates **visual robustness** and **neural alignment** in artificial agents. The benchmark includes a Unity-based foraging environment, held-out visual perturbations, standardized evaluation code, baseline agents, and paired visual-neural recordings from mice performing the same task. Agents are trained to perform a visually guided foraging task in a 3D Unity environment and are evaluated on their ability to generalize to a set of visual perturbations. Crucially, head-fixed mice perform a matched version of the same task in the same environment via a floating-ball setup, enabling comparisons grounded in shared visual input and task structure.

The benchmark enables three key evaluations. First, it measures robustness in an embodied setting, where perturbations affect perception within an active sensorimotor loop rather than static image classification. Second, it evaluates neural alignment as a property that emerges from task-driven learning, by assessing how well model representations predict neural activity without using neural data during training. Third, by evaluating both properties on the same models, the benchmark provides a framework to study whether robustness and neural alignment co-emerge or require distinct inductive biases. The perturbations target distinct aspects of visual processing, including contrast degradation, occlusion, and motion structure. To support reproducibility and community-driven extensions, we release the full Unity environment alongside the dataset and evaluation code. This enables future work to introduce new perturbations, design more complex task variants, and extend the benchmark to new evaluation settings.

2 Competition Overview

The Mouse vs. AI: Robust Foraging competition provides a unified benchmark to evaluate both visual robustness and neural alignment in artificial agents. Agents are trained to perform a visually guided navigation task and are evaluated on their ability to generalize across visual perturbations (Track 1) and on the ability of their internal representations to predict neural activity recorded in mice performing the same task (Track 2). By combining a shared behavioral task with a multimodal biological dataset, the benchmark enables systematic comparison between artificial and biological vision.

Participants train reinforcement learning agents to navigate to a target object based solely on egocentric visual input in a shared environment, where both artificial agents and mice perform the same task (Figure 1A). During training, agents are exposed to the environment under standard conditions and a single example perturbation (fog). After training, agents are evaluated on held-out perturbations to assess robustness (Track 1; Figure 1B), and their internal representations are evaluated for neural alignment using recorded mouse data (Track 2; Figure 1C).

Visual Foraging Task. The benchmark centers on a visually guided navigation task in which agents and mice must locate and approach a target object based solely on visual input. At each timestep, the agent receives an egocentric grayscale image and selects actions to navigate through the environment. In each trial, a gray target object appears at a fixed distance with a random horizontal offset (uniformly distributed within $\pm 30^\circ$). The agent must reach the target within five seconds for a trial to be counted as a *success*; otherwise, the trial ends in *failure*.

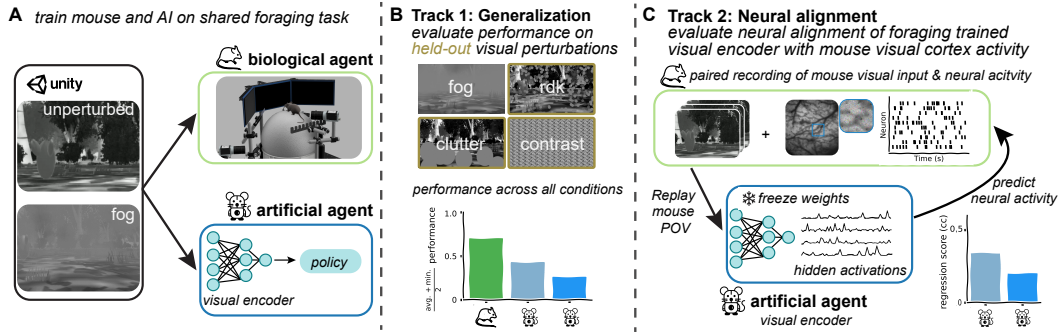


Figure 1: **Overview of the Mouse vs AI benchmark.** **A.** Mice and agents perform the same visually guided foraging task. **B.** Track 1 evaluates generalization across visual perturbations using success rate metrics. **C.** Track 2 evaluates neural alignment by predicting mouse neural activity from model representations.

The task is designed to be simple in objective but perceptually challenging. The target is embedded in a dynamic, visually complex environment with naturalistic textures, lighting, and motion, making it nontrivial to detect and track. Both mice and artificial agents perform the same task under matched conditions, enabling direct comparison between biological and artificial behavior. **Track 1: Visual Robustness.** Track 1 evaluates each agent’s success rate under held-out visual perturbations. These perturbations introduce structured changes to the visual input, requiring agents to generalize beyond training conditions.

The perturbations were designed to target distinct components of visual processing relevant for navigation. **Fog** reduces contrast and obscures distal landmarks, challenging long-range spatial perception. **Clutter** occludes large portions of the scene, requiring segmentation and integration beyond local visual cues. **Random dot kinematogram (RDK) overlays** disrupt global motion structure, interfering with optic flow processing. **Contrast-modulated noise** introduces structured visual interference that cannot be easily resolved by simple local filters. Together, these perturbations probe complementary aspects of visual processing, enabling a more fine-grained assessment of robustness than uniform noise or corruption benchmarks.

The goal is to develop agents that maintain task performance under distributional shifts in a behaviorally grounded setting.

Track 2: Neural Alignment. Track 2 evaluates the similarity between model representations and neural activity recorded from mouse visual cortex. For each trained model, visual input from the mouse experiments is passed through the model, and a linear readout is trained to predict recorded neural responses. Models are trained exclusively on the foraging task and do not use neural data during training; neural alignment is evaluated post hoc from fixed model representations.

3 Benchmark resources and data

The benchmark provides two complementary resources: a Unity-based environment for training and evaluating artificial agents (Track 1), and a paired visual–neural dataset recorded from mice performing the same task (Track 2). All resources are publicly available at <https://robustforaging.github.io/>.

Unity Environment (Track 1). Participants train agents via online reinforcement learning in a Unity-based simulator built on the ML-Agents framework. No pre-collected training dataset is provided; agents learn solely through interaction with the environment.

At each timestep, the agent receives a grayscale egocentric visual observation and outputs a continuous action controlling forward/backward translation, lateral translation, and rotation. The task is rewarded sparsely based on successful target acquisition.

We provide multiple Unity builds for training and validation under different visual conditions, including normal trials and a single example perturbation (fog). During evaluation, each model

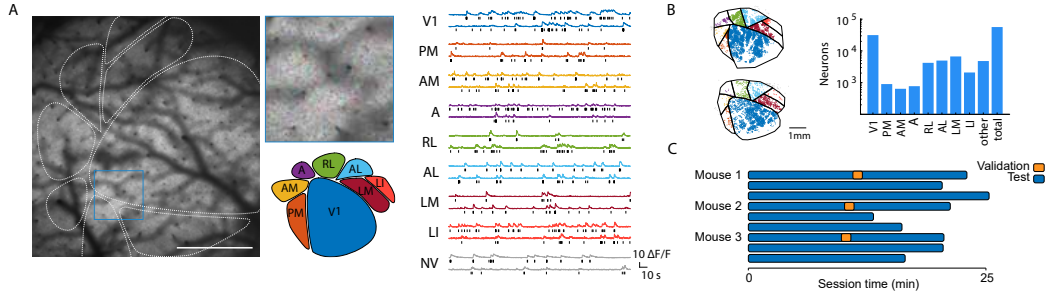


Figure 2: **Track two dataset: mesoscale neural recordings from visual cortex paired with stimulus frames during active VR navigation.** **A.** Raw mesoscale D2P recordings. Left, example imaging field of view (FOV) with overlay of cortical visual areas and segmented cell masks. Scale bar, 1 mm. Right, example calcium traces colored by visual area with inferred spike trains (black). Areas include nonvisual (NV) neurons outside visual cortex. **B.** Neuron counts. Left, all neurons from two example sessions and separate mice. Right, total neuron counts by visual area in the dataset. **C.** Recording length for all sessions and mice, with overlay of small validation set from competition.

is tested on 100 trials per condition across five visual settings: normal, fog, and three held-out perturbations (clutter, random dot kinematograms, and contrast-modulated noise). As a biological baseline, we provide trial-averaged success rates from $n = 3$ expert mice performing the same task. The full Unity codebase is released, enabling the community to introduce new perturbations and task variants beyond the initial competition scope.

124

Neural Dataset (Track 2). While mice performed the same foraging task in virtual reality, population activity was recorded using two-photon calcium imaging from layer 2/3 neurons in primary visual cortex (V1) and multiple higher visual areas. An overview of the dataset is shown in Figure 2, including example recordings across visual areas (Figure 2A), neuron counts (Figure 2B), and recording durations across sessions (Figure 2C). See Schneider et al. [2025] for recording details.

Preprocessing. Calcium signals were motion-corrected, segmented, and converted to inferred spike trains using standard pipelines [Pachitariu et al., 2017, Pnevmatikakis et al., 2013]. Spike trains were binned into 100 ms intervals and temporally aligned with visual input, accounting for neural response latency. A quality-control procedure was applied to retain a subset of reliably inferred neurons. Specifically, neurons with very low activity or weak visual responsiveness were excluded based on firing rate and predictive reliability criteria. This filtering reduces noise in the alignment scores by excluding neurons that do not exhibit consistent visually driven responses.

Dataset format. Each sample pairs a visual frame $S_t \in \mathbb{R}^{86 \times 155}$ with a neural activity vector $N_t \in \mathbb{N}^n$ representing binned spike counts:

$$\{(S_t, N_t)\}_{t=1}^T. \quad (1)$$

Evaluation protocol. Neural data are used exclusively for post hoc evaluation of model representations. During the competition, a small subset of recordings (approximately 1 minute per session across 3 sessions) was provided for validation of the evaluation pipeline. Models were trained solely on the foraging task without access to neural data.

For evaluation, visual input from the mouse experiments is passed through each trained model, and a linear readout is trained to predict recorded neural responses from fixed model activations. Alignment is quantified as the correlation between predicted and observed neural activity.

Dataset summary. The competition evaluation set consisted of 3 sessions from 3 mice (61 minutes total), comprising 19,746 neurons across up to 8 visual areas. The public release expands this dataset to 9 sessions from 3 mice, totaling approximately 178 minutes of recording and more than 50,000 neurons across multiple visual areas.

Code and data availability. The full benchmark, including Unity environments, evaluation pipelines for both tracks, and dataset access, is available at robustforaging.github.io. Dataset documentation

Model	Track 1	Track 2
HCMUS_TheFangs	0.9540	0.1547
alluding123	0.7570	0.1360
sazeglio	0.7360	0.1280
CNN Baseline	0.6150	0.1367
ResNet-18 Baseline	0.6260	0.1269
Xu2023 Baseline	0.2520	0.1305

Table 1: Performance of the top models and baselines on both tracks.

152 follows the Croissant machine-readable format as required by the NeurIPS Evaluations and Datasets
153 Track.

154 4 Competition Evaluation

155 Each track uses a distinct evaluation metric aligned with its objective. Final scores are computed
156 by the organizers using centralized infrastructure on held-out test data. Participants submit trained
157 models and inference code via a containerized environment to ensure consistent and reproducible
158 evaluation.

159 **Track 1: Visual Robustness.** Robustness is evaluated based on agent performance across five visual
160 conditions: normal and fog environments (available during training), and three held-out perturbations.
161 A trial is counted as successful if the agent reaches the target within the 5-second time limit.

162 Each agent is evaluated over 100 trials per condition using randomized seeds. Performance is
163 summarized using two metrics: the **Average Success Rate (ASR)**, defined as the mean success rate
164 across all conditions, and the **Minimum Success Rate (MSR)**, defined as the lowest success rate
165 across conditions. The final Track 1 score is computed as the mean of ASR and MSR, capturing both
166 overall performance and worst-case generalization.

167 **Track 2: Neural Alignment.** Neural alignment is evaluated by measuring how well model represen-
168 tations predict neural activity recorded from mice performing the same task. Visual input sequences
169 recorded during mouse behavior are replayed through the trained model with frozen weights, and
170 activations are extracted from each layer.

171 For each layer, a linear readout is trained to predict recorded neural responses from the model
172 activations (after dimensionality reduction). Formally,

$$\hat{N}_t = W M_t + b, \quad (2)$$

173 where M_t denotes the model activations at time t , and W and b are fitted parameters.

174 Performance is quantified as the mean Pearson correlation between predicted and observed responses
175 across neurons:

$$\rho = \frac{1}{n} \sum_{i=1}^n \text{corr}(\hat{N}_{t,i}, N_{t,i}). \quad (3)$$

176 The final alignment score is defined as the best-performing layer:

$$\text{Score}_2 = \max_l \rho^{(l)}. \quad (4)$$

177 Further implementation details and extended analysis of the evaluation procedure are provided in
178 Schneider et al. [2025].

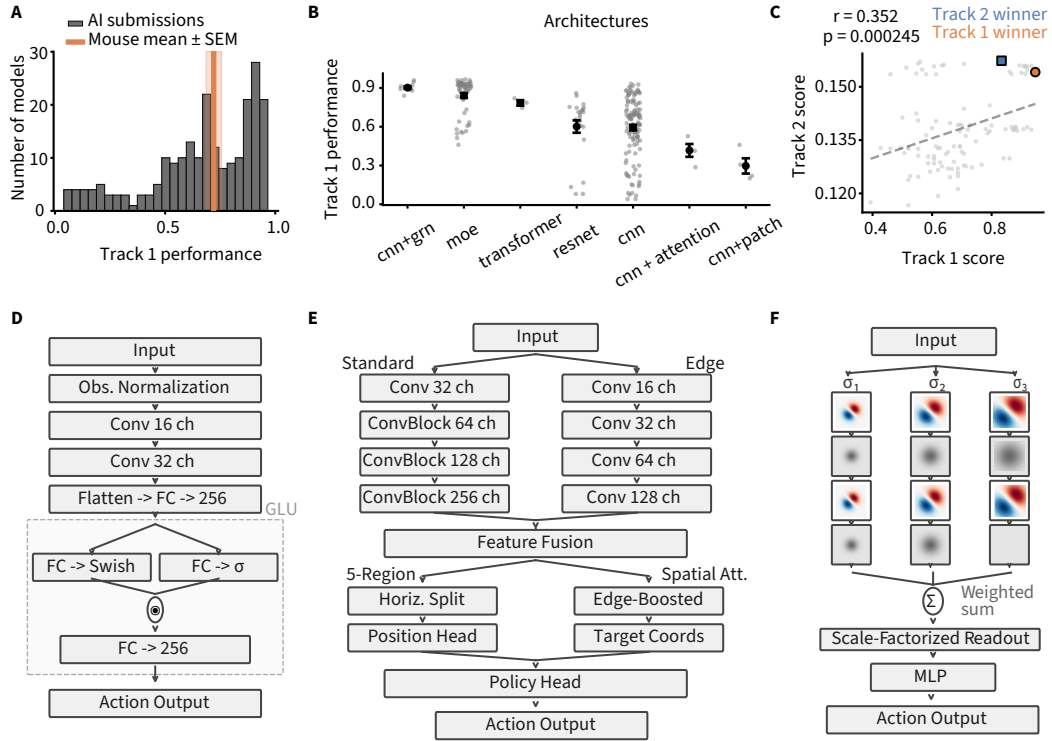


Figure 3: Performance, aggregate trends, and architectures of top models. **A.** Distribution of Track 1 (held-out) performance across all submissions, with the mean $\pm$ SEM of mouse performance shown for comparison. Model performance varies widely, highlighting the difficulty of achieving robust generalization under visual perturbations. **B.** Track 1 performance grouped by architectural family. Performance differences across architectures indicate that robustness depends more on specific inductive biases than on architecture class alone. **C.** Relationship between Track 1 (robustness) and Track 2 (neural alignment). A weak positive correlation ($r = 0.35$) indicates that the two objectives are only partially aligned, with models specializing along either axis or achieving moderate performance on both. **D-F** Architectures of the top three submissions. **(D)** First-place model (HCMUS_TheFangs): a lightweight CNN with observation normalization and GLU gating. **(E)** Second-place model (alluding123): a dual-branch architecture combining local and large-kernel feature extraction with spatial attention. **(F)** Third-place model (sazeglio): a scale-equivariant network with a factorized readout operating across spatial scales.

5 Baseline Models

To lower the barrier to entry, we provide three baseline agents trained using Proximal Policy Optimization (PPO) within the ML-Agents framework.

All models share a common architecture consisting of a visual encoder followed by a two-layer multilayer perceptron (MLP) policy head that maps visual features to continuous actions. The baselines differ only in the visual encoder:

- **Simple CNN:** a lightweight convolutional network with three convolutional layers.
- **ResNet-18:** a deeper convolutional architecture with residual connections [He et al., 2015].
- **Xu2023:** a model architecture adapted from Xu et al. [2023], previously shown to predict visual cortical activity in freely moving mice.

All agents are trained from scratch under identical PPO hyperparameters using the normal and fog environments. These baselines provide reference points for both robustness and neural alignment performance.

6 Results

6.1 Participation and overall performance

During the four-month submission period, 22 teams submitted a total of 290 models. Of these, 114 submissions outperformed the best baseline model on Track 1. Table 1 summarizes the performance of the top three models alongside the provided baselines.

Model performance varies substantially across submissions (Fig. 3A), ranging from near-random behavior to high success rates approaching expert-level mouse performance. This wide distribution highlights the difficulty of achieving robust generalization under visual perturbations. The winning model (HCMUS_TheFangs) outperforms the next-best submission by a large margin on Track 1, emphasizing that strong robustness emerges only under carefully tuned architectural and training choices.

6.2 Aggregate trends across submissions

We next analyze global trends across all submitted models to understand how robustness and neural alignment relate.

Figure 3B shows Track 1 performance grouped by visual encoder architecture. While performance varies widely within each category, several architectures incorporating additional structure—such as recurrence, mixture-of-experts, or transformer-based encoders—tend to achieve higher robustness. In contrast, CNN-based models span the full range of performance, indicating that architectural class alone does not determine robustness.

Figure 3C shows the relationship between Track 1 (held-out robustness) and Track 2 (neural alignment). We observe a weak positive correlation between the two metrics ($r = 0.35$), indicating that models that generalize better under perturbations tend to exhibit slightly higher neural alignment, but that the two objectives are not strongly coupled. Several models specialize along one axis, achieving high robustness but low alignment, or vice versa, while a smaller subset achieves moderate performance on both tasks. This distribution suggests that robustness and neural alignment occupy partially overlapping but distinct regions of the model design space.

To further investigate factors influencing performance, we examine relationships between the performance on the two tracks and basic model and training properties (Appendix Figure 5). While robustness shows little dependence on model size, neural alignment increases with model depth and, to a lesser extent, parameter count. This indicates that increasing model capacity alone does not improve robustness, reinforcing that inductive biases rather than scale drive performance under distributional shift. In contrast, the strong dependence of neural alignment on depth suggests that hierarchical representations are critical for capturing the structure of neural responses.

We also observe a non-monotonic relationship between training duration and performance, with intermediate training steps yielding higher robustness and alignment than longer training runs (Appendix Figure 5). This non-monotonic trend resembles an inverted-U pattern, suggesting that extended training may lead to overfitting to the training conditions and reduced generalization under perturbations. This pattern is consistent with the winning team’s observation that early stopping can improve generalization.

Finally, recurrent architectures tend to achieve higher robustness under held-out conditions (Appendix Figure 6), suggesting that temporal integration may provide an additional mechanism for resilience to visual perturbations.

While our primary analysis focuses on Track 1 performance, we note that the top-performing Track 2 models consistently rely on deeper architectures, emphasizing representational capacity over simplicity. This contrast further supports the observation that robustness and neural alignment favor different architectural regimes.

6.3 Winning solutions

First Place: HCMUS_TheFangs. The winning team achieved top performance by adopting distinct architectures tailored to the two evaluation objectives (Fig. 3D). For Track 1, they used a lightweight convolutional model augmented with Gated Linear Units (GLU) and observation normalization.

Despite its simplicity, this model achieved the highest robustness under held-out perturbations. Ablation studies revealed that observation normalization contributed the largest performance gain, while increasing architectural complexity—including deeper residual networks, recurrent modules, or data augmentation—often degraded robustness.

In contrast, their Track 2 solution relied on a substantially deeper convolutional architecture with increased capacity, enabling improved prediction of neural activity. This separation highlights a key insight: robustness and neural alignment favor different regions of the model design space. While shallow models generalize better under distributional shift, deeper architectures are required to capture the structure of neural responses.

Training dynamics further reinforced this distinction. Performance was highly sensitive to training duration, with optimal results achieved at intermediate checkpoints, while extended training often reduced both robustness and alignment. Together, these findings suggest that both architectural simplicity and carefully tuned training procedures are critical for robust generalization.

Second Place: alluding123. The second-place solution introduces a dual-branch visual encoder designed to capture complementary spatial features (Fig. 3E). One branch focuses on fine-grained local structure using standard convolutional layers, while a parallel branch employs large-kernel convolutions to capture broader contextual information and improve sensitivity to targets at the periphery of the visual field. The outputs of both branches are fused into a unified representation, enabling the model to integrate information across spatial scales.

In addition, the architecture incorporates explicit spatial reasoning mechanisms, including a position detection head and an edge-aware attention module that redistributes focus toward peripheral regions. This design addresses a key challenge of the task, namely detecting targets at extreme positions under perturbations. By combining multi-scale feature extraction with spatially structured attention, the model achieves strong robustness without relying on excessive model depth.

Third Place: sazeglio. The third-place solution explores a fundamentally different approach based on scale-equivariant convolutional networks (Fig. 3F). By operating in a joint space-scale representation, the model preserves information across multiple spatial scales, enabling consistent detection of objects under changes in size and visual appearance.

A key component of this architecture is a factorized readout that separately models spatial and scale-dependent contributions, allowing the network to retain scale-specific information rather than collapsing it through pooling. This design achieves substantial parameter efficiency while maintaining competitive performance, suggesting that structured representations can provide an alternative to scaling model size. Extensive visual augmentations during training further improved robustness to diverse perturbations.

Together, these solutions highlight that robust behavior can be achieved through diverse architectural strategies, ranging from minimal models with carefully chosen inductive biases to more structured representations that explicitly encode spatial and scale information.

7 Discussion

The Mouse vs. AI competition attracted 22 teams and 290 model submissions, providing a diverse testbed for analyzing the relationship between visual robustness and neural alignment. By combining a shared behavioral task with neural recordings, the benchmark enables a direct comparison between generalization under perturbations and the biological plausibility of learned representations.

Across submissions, three main findings emerge. First, robustness and neural alignment are only weakly coupled. Models that generalize better under perturbations tend to exhibit modestly higher neural alignment, the relationship is modest, and many models specialize along one axis. This suggests that optimizing for robust behavior does not automatically yield brain-like representations, and vice versa, although a subset of models achieves moderate performance on both objectives.

Second, architectural design and model capacity play distinct roles across tasks. Deeper models tend to achieve higher neural alignment, consistent with the need for hierarchical representations to predict neural responses. In contrast, robustness does not systematically improve with model size, and simpler architectures often perform better under distributional shift. Notably, the top-performing

293 solutions rely on targeted inductive biases (e.g., normalization, gating, structured feature extraction)
294 rather than generic model scaling, suggesting that task-specific architectural design can be more
295 important than increasing capacity.

296 Third, training dynamics play a critical role. Both the winning team’s analysis and aggregate trends
297 across submissions indicate that performance does not improve monotonically with training duration.
298 Models trained for intermediate durations frequently outperform longer-trained models, suggesting
299 that early stopping and checkpoint selection may be more important than total training compute for
300 achieving both robustness and neural alignment.

301 Together, these findings indicate that robustness and neural alignment favor different regions of the
302 model design space, and that no single architectural or training strategy optimizes both objectives
303 simultaneously. Rather than identifying a single “best” model, the benchmark reveals trade-offs that
304 can inform future model development.

305 **Limitations.** Several limitations constrain the conclusions supported by this benchmark. The neural
306 dataset is drawn from a limited number of animals, restricting the assessment of biological variability.
307 The benchmark focuses on a single foraging task, and it remains an open question whether the
308 observed relationships generalize to other visually guided behaviors. The perturbation set covers a
309 limited number of conditions, and robustness conclusions are bounded by this scope.

310 On the modeling side, the aggregate analyses are observational and do not establish causal relation-
311 ships between architectural features and performance. Additionally, calcium imaging captures slower
312 dynamics than spiking activity, and the head-fixed virtual reality setup constrains the naturalistic
313 sensorimotor experience of the animal.

314 **Conclusion.** Most existing NeuroAI benchmarks evaluate neural alignment under passive viewing
315 conditions. In contrast, the Mouse vs. AI benchmark evaluates models in an active, behaviorally
316 grounded setting with dynamic and perturbed visual input. By jointly measuring robustness and neural
317 alignment within a shared task, the benchmark provides a more ecologically relevant framework
318 for studying the relationship between artificial and biological vision. We hope that this benchmark
319 facilitates the development of models that are both robust in real-world conditions and aligned with
320 neural representations.

References

- Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliki, Sharada P Mohanty, Florian Laurent, Marcel Salathé, Matthias Bethge, Yaodong Yu, et al. Adversarial vision challenge. In *The NeurIPS’18 Competition: From Machine Learning to Intelligent Conversations*, pages 129–153. Springer, 2020.
- Mozhou Gao, Chris H. Hugenholtz, Thomas A. Fox, Maja Kucharczyk, Thomas E. Barchyn, and Paul R. Nesbit. Weather constraints on global drone flyability. *Scientific Reports*, 11(1):12092, June 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-91325-w. URL <https://www.nature.com/articles/s41598-021-91325-w>. Publisher: Nature Publishing Group.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL <http://arxiv.org/abs/1512.03385>. arXiv:1512.03385 [cs].
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.
- Chengjie Lu, Tao Yue, Man Zhang, and Shaukat Ali. DeepQTest: Testing Autonomous Driving Systems with Reinforcement Learning and Real-world Weather Data, October 2023. URL <http://arxiv.org/abs/2310.05170>. arXiv:2310.05170 [cs].
- Marius Pachitariu, Carsen Stringer, Mario Dipoppa, Sylvia Schröder, L. Federico Rossi, Henry Dalgleish, Matteo Carandini, and Kenneth D. Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv*, page 061507, July 2017. doi: 10.1101/061507. URL <https://www.biorxiv.org/content/10.1101/061507v2>. Publisher: Cold Spring Harbor Laboratory Section: New Results.
- Eftychios A. Pnevmatikakis, Josh Merel, Ari Pakman, and Liam Paninski. Bayesian spike inference from calcium imaging data. *arXiv:1311.6864 [q-bio, stat]*, November 2013. URL <http://arxiv.org/abs/1311.6864>. arXiv: 1311.6864.
- Marius Schneider, Joe Canzano, Jing Peng, Yuchen Hou, Spencer LaVere Smith, and Michael Beyeler. Mouse vs. ai: A neuroethological benchmark for visual robustness and neural alignment. *ArXiv*, pages arXiv–2509, 2025.
- Martin Schrimpf, Jonas Kubilius, Alex Lee, N Ratan Murty, Robert Ajemian, and James J DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *Advances in Neural Information Processing Systems*, 31, 2018.
- Polina Turishcheva, Paul G. Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin F. Willeke, Mohammad Bashiri, Ruslan Baikulov, Yu Zhu, Lei Ma, Shan Yu, Tiejun Huang, Bryan M. Li, Wolf De Wulf, Nina Kudryashova, Matthias H. Hennig, Nathalie L. Rochefort, Arno Onken, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. Retrospective for the Dynamic Sensorium Competition for predicting large-scale mouse primary visual cortex activity from videos, July 2024. URL <http://arxiv.org/abs/2407.09100>. arXiv:2407.09100 [q-bio].
- Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A Survey on the Robustness of Computer Vision Models against Common Corruptions, September 2024. URL <http://arxiv.org/abs/2305.06024>. arXiv:2305.06024 [cs].
- Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The Sensorium competition on predicting large-scale mouse primary visual cortex activity, June 2022. URL <http://arxiv.org/abs/2206.08666>. arXiv:2206.08666 [q-bio].
- Aiwen Xu, Yuchen Hou, Cristopher M. Niell, and Michael Beyeler. Multimodal Deep Learning Model Unveils Behavioral Dynamics of V1 Activity in Freely Moving Mice. In *The Thirty-Seventh Annual Conference on Neural Information Processing Systems*, May 2023. doi: 10.1101/2023.05.30.542912. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/31a19921acd38cdf7a8c86ec032cef2d-Paper-Conference.pdf.

8 Acknowledgments

This work was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number R01-NS121919. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

MS: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Data Curation, Validation, Visualization, Writing - original draft, Writing - Review and Editing
JC: Conceptualization, Formal Analysis, Investigation, Methodology, Animal Experiments, Software, Data Curation, Validation, Visualization, Writing - original draft, Writing - Review and Editing
YH: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Writing - original draft, Writing - Review and Editing
JP: Investigation, Software, Validation, Methodology, Project Administration, Visualization, Writing - Review and Editing
AD, UK, PHP, CNT DSDM, NLPQ, HTK & SA: Winning Participants, Writing - original draft
SLS: Conceptualization, Methodology, Funding Acquisition, Supervision, Writing - Review and Editing
MB: Conceptualization, Methodology, Funding Acquisition, Supervision, Writing - Review and Editing

9 Supplementary Figures

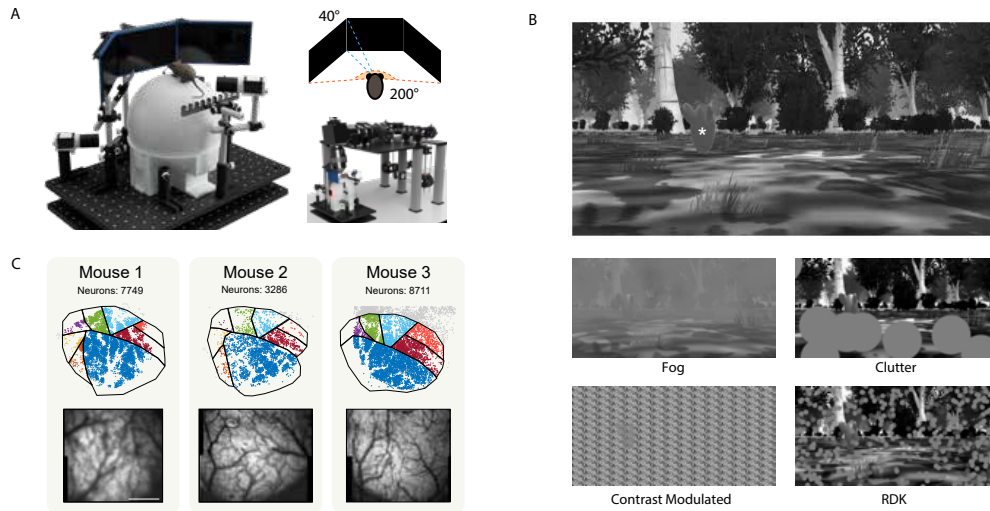


Figure 4: VR task and neural dataset details. **A.** VR behavior apparatus. Top right, visual field covered by screen array. Bottom right, VR apparatus deployed under D2P for imaging experiment. **B.** Task visual environment and perturbations. Top, normal stimulus. Bottom, perturbations. **C.** Example sessions from each mouse. Top, all segmented neurons from one recording session. Bottom, example D2P fields of view from same animals.

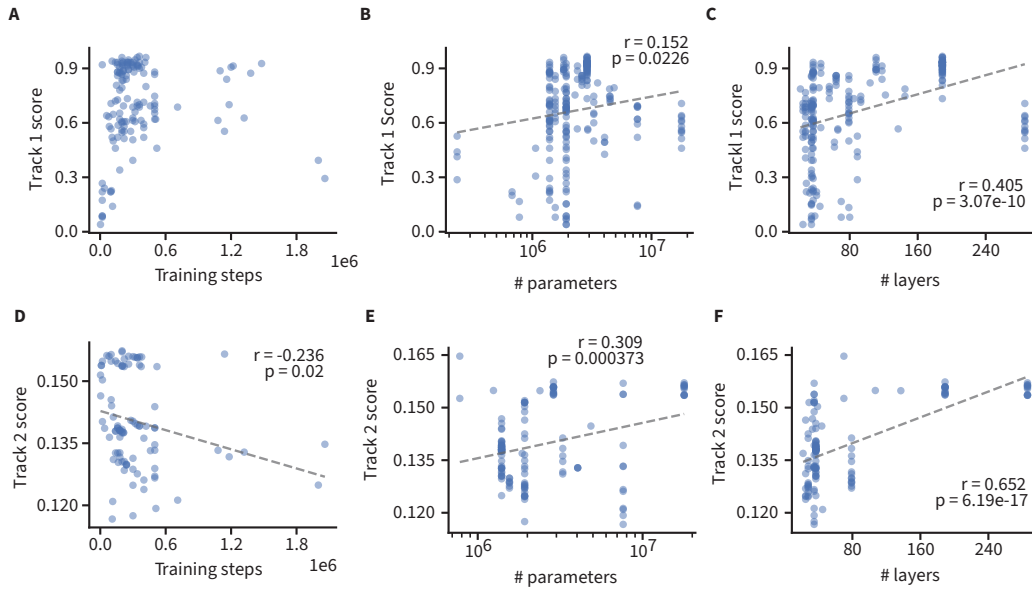


Figure 5: **Scaling relationships between model properties and performance.** (A–C) Relationship between Track 1 performance and (A) training steps, (B) number of parameters, and (C) model depth. Robustness shows little dependence on model size, but a modest increase with depth. (D–F) Relationship between Track 2 performance and (D) training steps, (E) number of parameters, and (F) model depth. Neural alignment increases strongly with model depth and moderately with parameter count, while showing a negative trend with extended training. Across both tracks, training duration exhibits a non-monotonic relationship with performance, suggesting that intermediate training durations yield better generalization than prolonged training.

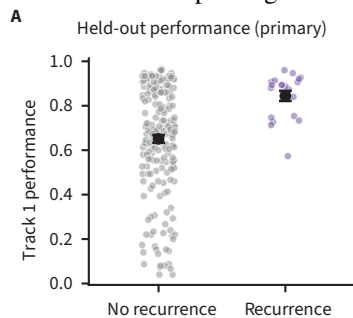


Figure 6: **Effect of recurrence on robustness.** A. Comparison of Track 1 held-out performance for models with and without recurrent components. Models incorporating recurrence achieve consistently higher robustness, suggesting that temporal integration provides an additional mechanism for improving generalization under visual perturbations.

10 Supplementary materials

10.1 Dataset

We release two versions of the neural dataset.

- **Preprocessed data:** Paired visual frames and inferred spike trains, ready for Track 2 evaluation. This is the dataset used by the benchmark evaluation code. Available at
- **Raw data:** Raw paired visual frames and inferred spikes without visually responsive cell filtering, for completeness.

Both are available at: <https://dataverse.harvard.edu/previewurl.xhtml?token=26e79441-e736-4805-ba7f-a7d00dbd3bb7>

399 **10.2 Croissant Metadata**

400 Machine-readable dataset documentation in Croissant format is available in the same via Harvard
401 Dataverse dataset as above: [https://dataverse.harvard.edu/previewurl.xhtml?token=](https://dataverse.harvard.edu/previewurl.xhtml?token=26e79441-e736-4805-ba7f-a7d00dbd3bb7)
402 [26e79441-e736-4805-ba7f-a7d00dbd3bb7](https://dataverse.harvard.edu/previewurl.xhtml?token=26e79441-e736-4805-ba7f-a7d00dbd3bb7)

403 **10.3 Code and Data Availability**

- 404 • Evaluation code and baselines: [https://github.com/SchneiderMarius/](https://github.com/SchneiderMarius/mouse-vs-ai-benchmark)
405 [mouse-vs-ai-benchmark](https://github.com/SchneiderMarius/mouse-vs-ai-benchmark)
- 406 • Neural dataset for Track 2 evaluation:
407 [https://dataverse.harvard.edu/previewurl.xhtml?token=](https://dataverse.harvard.edu/previewurl.xhtml?token=26e79441-e736-4805-ba7f-a7d00dbd3bb7)
408 [26e79441-e736-4805-ba7f-a7d00dbd3bb7](https://dataverse.harvard.edu/previewurl.xhtml?token=26e79441-e736-4805-ba7f-a7d00dbd3bb7)
- 409 • Unity environment: <https://github.com/SchneiderMarius/mouse-vs-ai-unity>
- 410 • Competition website: <https://robustforaging.github.io>

411 **Compute resources.** Baseline agents were trained on a single NVIDIA RTX 4090, with training
412 times ranging from 1-2 hours for lightweight models to 8+ hours for more complex architectures.
413 Track 1 evaluation runs inside the Unity engine and takes approximately 30 minutes per model
414 (100 trials $\times$ 5 conditions). Track 2 evaluation (linear readout fitting across all sessions) takes
415 approximately 15-30 minutes per model on a single GPU.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly state the paper’s contributions: a shared sensorimotor benchmark with two tracks (visual robustness and neural alignment), a Unity-based environment, a neural dataset of 178 minutes from 50,000+ neurons, and results from 22 competing teams. The scope and limitations of conclusions are described in section 7.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Section 7 includes an explicit limitations paragraph discussing the count of included animals, restriction to a single foraging task, limited perturbation set, observational (non-causal) nature of aggregate analyses, and constraints of calcium imaging and the head-fixed VR setup.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: This paper does not present theoretical results or proofs. The contributions are empirical: a benchmark, dataset, and competition results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Evaluation protocols for both tracks are described in detail in Sections 3 and 4, including the number of trials, metrics, and neural alignment procedure. Code and data are publicly released (Section 3, Supplementary 10.3).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All benchmark resources are publicly available at <https://robustforaging.github.io/>, including evaluation code and baselines (GitHub), the neural dataset (Harvard Dataverse), and the Unity environment (GitHub). Links are provided in Section 3 and Supplementary 10.3.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4 specifies the evaluation metrics, number of trials per condition, the five stimulus conditions, the linear readout training procedure, and the dimensionality reduction steps. Baseline training details (PPO, ML-Agents framework, shared hyperparameters) are described in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All aggregate values are reported either with mean $\pm$ SEM (Figure 3A-B, Figure 6A) or correlations with Pearson r and p -values (Figure 3C and Figure 5) where appropriate.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: This information was not collected from the 22 competing teams.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research involves animal experiments (mice) and machine learning benchmarking; no human subjects were used in data collection. Animal experiments follow standard neuroscience protocols as referenced in Schneider et al. [2025]. The benchmark does not pose ethical risks associated with harmful applications.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is basic neuroscience and machine learning benchmarking research. The benchmark is designed to advance understanding of visual robustness and neural alignment, and we are aware of no direct path this has to negative impacts on society.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released assets (Unity environment, neural dataset, evaluation code) pose no high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Third-party assets used in this work are properly cited, including the Unity engine and ML-Agents framework, as well as standard neuroscience tools such as Suite2p (Pachitariu et al., 2017), and the CNMF spike inference pipeline (Pnevmatikakis et al., 2013). The Xu2023 baseline architecture is attributed to Xu et al. [2023].

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new neural dataset, Unity benchmark environments for development and testing, and code for using the above assets to train and evaluate models. All released assets are well documented, see section 10 for details.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. Competition participants submitted models voluntarily and are credited as authors or acknowledged accordingly.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects were involved in this research. Animal experiments were conducted following IACUC approved protocols as described in the referenced companion paper (Schneider et al., 2025).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a core methodological component in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.