FACTORIZATION REGRET MEDIATES COMPOSITIONAL GENERALIZATION IN LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Are there still barriers to generalization once all relevant variables are known? We consider the challenge of generalizing to a novel combination of task-relevant latent variables in the *Cognitive Gridworld*, a stationary Partially Observable Markov Decision Process where observations are generated by latent variables with parametric interactions. We establish *Factorization Regret* as an information-theoretic quantity measuring the contribution of latent variable interactions to task performance. Using this metric, we explain the sub-optimality of Echo State Networks and uncover a failure mode whereby confidence becomes decoupled from accuracy. This suggests that utilizing interactions between relevant variables is a non-trivial capability. We next introduce *Representation Classification Chains* to show that latent variable interactions and how to infer the variables' values can be learned together through Reinforcement Learning. Solving this variational inference problem enables generalization to novel combinations of relevant variables. Lastly, we demonstrate that once interactions are learned, imagined experience is sufficient for zero-shot control. In summary, we develop and motivate a minimal, theoretically grounded setting for research into generalization from experience.

1 INTRODUCTION

As Machine Learning continues to solve increasingly challenging problems, unreliability remains a significant barrier. One central obstacle is that what *can* be learned depends crucially on what *has* been learned. For instance, inferring *time of day* from *ambient light* becomes much more reliable after learning about *seasons*. A combinatorial explosion can emerge from the interactions between variables, but typically only a few are relevant to a given task. Thus, the world may be compressed into something much more manageable, insofar as one's goal is concerned (Tishby et al., 2000; Roy et al., 2005). Yet, despite needing only to learn that which is immediately relevant for survival, animals develop an intuition about the natural world that is far more reliable than reasoning models (Tenenbaum, 2018; Philip & Hemang, 2024).

To explain this intuition, it helps to differentiate *meaningful* intelligence from that which is trivial. Chollet (2019) defines intelligence as 'skill-acquisition efficiency', emphasizing the capacity to build upon prior knowledge. This relies on compositional generalization, the capacity to generalize to an unseen combination of elements. In this paper, we consider compositional generalization with latent variables as the elemental units. The mutual information among three or more random variables is known as *Interaction Information* (Watanabe, 1960). Interaction Information has been used to understand sensory modality integration (Proca et al., 2024) and causal inference (Ghassami & Kiyavash, 2017). In this paper we connect Interaction Information with the notion of *Semantic Information* as the subset of information meaningful for survival (Kolchinsky & Wolpert, 2018). To exemplify their relationship, consider an agent with two actions (i.e., *causal interventions*), where reward is determined by an *exclusive or* (XOR) between the selected action and a binary variable. XOR demonstrates Interaction Information in its purest form (Rosas et al., 2019): Action and reward are independent unless the agent knows the value of the binary variable, so information about its value is semantic; scrambling it, as in Kolchinsky & Wolpert (2018), leaves reward at chance. We consider such variables *contextual* since they mediate dependencies between action, observations, and reward without being directly causal. Here, we introduce **Factorization Regret (FR)**: the contribution of the latent variable interactions to inference of the optimal action. More generally, FR emerges when conditioning on observables changes the mutual information between contextual vari-

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

ables and the optimal action. Conceptually, FR measures how much more reliable a policy becomes after accounting for latent variable interactions.

We formalize *meaningful* intelligence via Information Theory and Reinforcement Learning as meta-learning from reward how to perform inference in environments with compositional latent structure. We support this view with four contributions: **(1)** A family of Partially Observable Markov Decision Processes (POMDPs) in which Factorization Regret can be manipulated (subsection 4.1). **(2)** Factorization Regret's account of the performance gap between Fully Trained Recurrent Neural Networks (RNNs) and Echo State Networks (ESNs) (subsection 4.2). **(3)** An architecture which explicates how latent variable interactions can be learned from naturalistic feedback (subsection 4.3). **(4)** An account of how learning interactions enables zero-shot control in novel settings (subsection 4.4).

2 Related Works

2.1 Compositional Generalization

Animals learn to interact with the environment via the reinforcement of actions that maximize *reward* given the state of the world (Skinner, 1953; Bellman, 1957; Sutton & Barto, 1998; Russell & Norvig, 2003). By learning how different elements of the world interact with each other, an animal may be able to generalize to a novel environment (Hayes et al., 2001). Generalizing to a novel combination of elements is known as *compositional generalization* (Kay et al., 2024; Lampinen et al., 2025), and occurs naturally in recurrent networks trained to predict complex sensory inputs (Boeshertz & Clopath, 2025). For instance, the Hippocampus (HPC) has been modeled as a network that learns a compositional grid of sensory predictions (Whittington et al., 2020). Navigating this grid is considered an "internal action" that encodes object relationships and facilitates prediction. Similarly, models of decision-making consider sequential inference akin to navigation towards a "decision boundary" (Shadlen & Newsome, 2001; Gold & Shadlen, 2007; Kiani & Shadlen, 2009; Mante et al., 2013; Brunton et al., 2013; de Lafuente et al., 2015; Kang et al., 2017; Pereira et al., 2021; Safavi & Dayan, 2022; 2025). Treating the internal 'commitments' of a decision-making process as elemental units could explain how animals generalize to novel combinations of task-relevant outputs. The Cognitive Gridworld is intended as a minimal setting for studying such generalization, complementary to large-scale benchmarks such as ARC-AGI (Chollet et al., 2024), without the need for large-scale agents such as Gato (Reed et al., 2022) and Dreamer (Hafner et al., 2025).

2.2 Partially Observable Markov Decision Processes

Animals often learn without concrete examples or explicit symbols; for instance, riding a bike becomes automatic without a full understanding of its mechanics. Such abstract knowledge is often implicit—a 'gut-feeling' that is effortlessly utilized, but hard to describe, which guides action without access to the exact state of the world. Decision-making under this kind of uncertainty is naturally modeled via POMDPs (Monahan, 1982; Kaelbling et al., 1998) that utilize a *belief-state*, a probability distribution over all possibilities, updated by observation or active exploration (Hu et al., 2018; Karni et al., 2025). This belief-state is often intractable, requiring that relevant information be compressed into an *approximate belief-state* (Åström, 1965; Roy et al., 2005). For instance, clever use of memory could store the minimal statistics sufficient for *marginal* belief updating. Dopamine, which encodes belief-like reward prediction errors, could embed such statistics into the low-dimensional manifold that cortical dynamics traverse (Hollerman & Schultz, 1998; Mante et al., 2013; Hennig et al., 2023; Churchland & Shenoy, 2024). By using observations that lack spatiotemporal structure, we can isolate dynamics that arise purely from the intractability of the belief-state.

2.3 Meta-learning

Animals cluster observations by independent *latent causes* (Gershman & Niv, 2010; Dubreuil et al., 2022) via mechanisms of the brain such as attention and computation through dynamics (Niv et al., 2015; Vyas et al., 2020; Versteeg et al., 2025). Together, these form a traversable *cognitive map*—an abstract space of task-relevant variables (Tolman, 1948; Behrens et al., 2018; Garvert et al., 2023). This process can be described as meta-learning, wherein agents learn to do inference (Santoro et al., 2016; Wang et al., 2018; Yang et al., 2019; Hospedales et al., 2022; Turner & Barak, 2023; Lake &

Baroni, 2023; Bowler et al., 2025; Lan et al., 2025). Meta-learning owes its flexibility to the compositionality of the underlying generative process. Utilizing this compositionality, Achille et al. (2018) developed VASE, an algorithm that generates novel data augmentations by recombining learned generative factors. Similar approaches have been useful in Reinforcement Learning (RL) settings for few-shot, zero-shot, in-context and offline learning (Zintgraf et al., 2019; Rakelly et al., 2019; Chen et al., 2021; Laskin et al., 2022; Ajay et al., 2022). For all its flexibility, meta-learning is formally defined in a world that is fixed but unknown (Ortega et al., 2019). Yet, the complexity of meta-learning tasks hinders theoretical investigation of how dynamics *learn to meta-learn*. Here, we design meta-learning tasks with a flexible generative process that remains analytically tractable.

3 Problem Setup

We assume that only a few latent variables are relevant to a specific goal. We refer to the actual value of a latent variable as its **realization**, governed by a random process where each variable, in a **context** of C relevant variables, takes a value from R possible realizations:

$$\mathbf{r} \overset{\text{i.i.d.}}{\sim} U(\{1,\dots,R\})^C, \quad \text{where } r_c \in \mathbf{r} \text{ is the realization of variable } c.$$

The agent observes the world through a d_o-dimensional observation space $\mathcal{O}$ (Figure 1). At time t, the probability of each independent observable $o^i \in \{0,1\}$ within observation $\mathbf{o}_t$ depends on realizations $\mathbf{r}$ and a likelihood parameterized by interactions Z. When $C = 1$, variables only interact with themselves. For all $C \geq 2$ there are $|Z| = d_o \times C \times (C-1)$ pairwise interactions between latent variables. $\mathbf{z}^i \in Z$ are the interactions relevant to o^i. We let $\ell_{\mathbf{z}^i}(\mathbf{r})$ denote the likelihood $P_{\mathbf{z}^i}(o^i = 1 \mid \mathbf{r})$, forming a tensor of order C with R entries along each axis. Construction of $\ell_{\mathbf{z}^i}(\mathbf{r})$ is described in detail in subsection A.2; a brief description follows.

Each pairwise interaction $z^i_{cc'}$ (the modulation from variable c' onto c) shifts the phase of a sinusoidal base to form vectors $\mathbf{v}^i_{cc'}(r)$, functionally "rolling" values of $\ell_{\mathbf{z}^i}(\mathbf{r})$ along axis c. When $C = 1$, the likelihoods are given by vector $\ell_{\mathbf{z}^i} = \sigma(\mathbf{v}^i)$, where σ denotes element-wise sigmoid. For all $C \geq 2$, broadcasting operator $\mathcal{B}_{cc'}$ entangles r_c and $r_{c'}$ by mapping the $R \times R$ outer product $\mathbf{v}^i_{cc'}(\mathbf{v}^i_{c'c})^\top$ onto the C-dimensional likelihood tensor $\ell_{\mathbf{z}^i}$:

$$\ell_{\mathbf{z}^i} = \sigma\left(\sum_{c<c'} \mathcal{B}_{cc'}\left(\mathbf{v}^i_{cc'}(\mathbf{v}^i_{c'c})^\top\right)\right) \quad C \geq 2$$

An episode E is defined by the tuple $(\mathbf{r}, g, Z, \mathbf{o}_{1:T})$, comprising the realizations, a random goal index $g \sim U(\{1,\dots,C\})$, the interactions, and a trajectory of T observations. The agent's objective is to infer the realization of goal variable r_g and select optimal action $a^\star = r_g$ from $a \in \{1,\dots,R\}$. Optimal inference of r_g can be achieved with sequential Bayesian updating:

$$P_Z(\mathbf{r} \mid \mathbf{o}_{1:t}) \propto P_Z(\mathbf{o}_t \mid \mathbf{r})\, P_Z(\mathbf{r} \mid \mathbf{o}_{1:t-1}),$$

where the joint likelihood of the independent observables follows a product of Bernoullis:

$$P_Z(\mathbf{o}_t \mid \mathbf{r}) = \prod_{i=1}^{d_o} \ell_{\mathbf{z}^i}(\mathbf{r})^{o^i_t}(1 - \ell_{\mathbf{z}^i}(\mathbf{r}))^{1-o^i_t}.$$

4 Results

4.1 Factorization Regret quantifies the cost of Naive Bayes

A natural baseline is a Bayesian observer that updates marginal beliefs (B_{tc}) independently. The likelihood for this *Naive* observer is acquired by marginalizing the joint likelihood (Figure 2a). We quantify the contribution of latent variable interactions to task performance at time t with a Kullback-Leibler divergence ($\mathcal{D}_{\text{KL}}$) between the Naive and Optimal Joint Bayesian observers' marginal beliefs over the goal variable, which we refer to as Factorization Regret (derived in subsection B.1).

$$\text{Factorization Regret}_t = \mathcal{D}_{\text{KL}}(B^{\text{Joint}}_{tg} \| B^{\text{Naive}}_{tg})$$

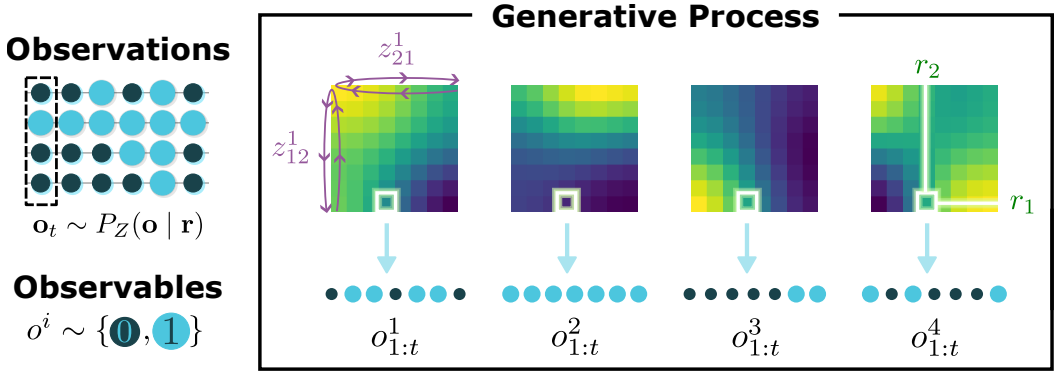

Figure 1: **Environment schematic for $C = 2$.** Observations $\mathbf{o}_t$ are generated stochastically. Variable interactions Z parameterize likelihood $P_Z(\mathbf{o} \mid \mathbf{r})$ and variable realizations (r_1, r_2) fix the probability of sampling each observable o^i.

FR quantifies the accumulating cost of scrambling Interaction Information on every step.

Figure 2b shows that average performance (calculated as $\langle [\arg\max_r (B_{tgr}) = a^\star] \rangle_E$ using Iverson bracket notation) improves with time and declines with the number of interacting variables, for both observers. However, the relative accuracy of joint and naive observers grows rapidly with interactions, along with FR. In summary, we have a family of POMDPs where the gap between Joint and Naive Bayes can be manipulated and attributed entirely to latent variable interactions.

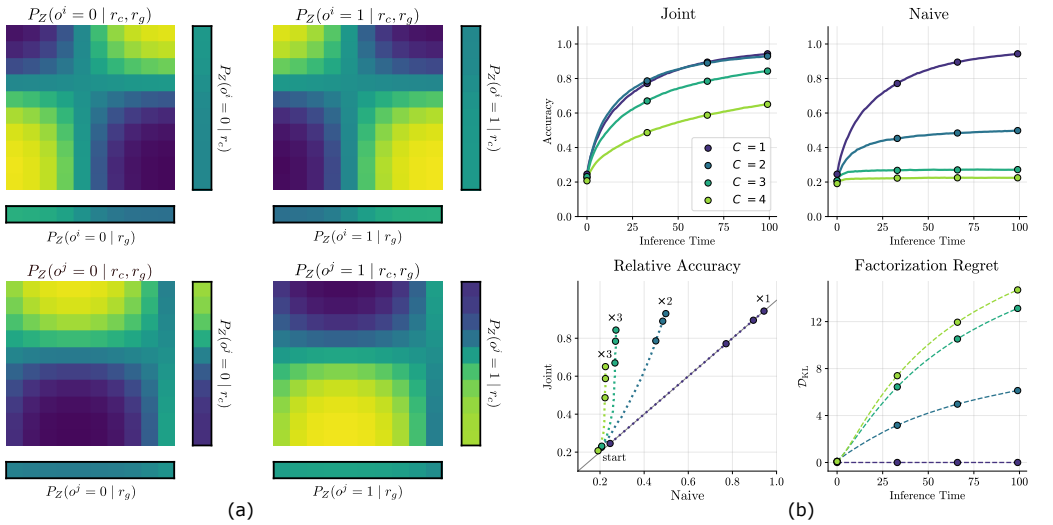

Figure 2: **The cost of Naive Bayes grows with time and interactions. (a)** Example Joint (matrices) and marginalized (vectors) likelihoods. **(b)** Top: Accuracy of Joint (left) and Naive (right) Bayes across varying context sizes. Bottom: Relative accuracy (left) and Factorization Regret (right). Circles mark four equidistant reference time-points throughout inference.

4.2 Experiment 1: The Cognitive Gridworld requires flexible neural dynamics

To demonstrate the practical consequences of Factorization Regret, we compare Fully Trained RNNs and ESNs, commonly used as models of neural dynamics (Mante et al., 2013; Wang et al., 2018; Yang et al., 2019). Echo State dynamics provide random temporal features for a readout to learn an arbitrary function of past inputs. Yet, they suffer from tradeoffs between discrimination versus generalization, memory versus nonlinearity, and recall versus precision (Alvarez, 2002; Dambre et al., 2012; Barak et al., 2013; Inubushi & Yoshimura, 2017; Tiňo, 2020; Tiňo et al., 2024). We find that Naive Bayes predicts the precise sub-optimal performance achieved by ESNs.

Interaction differentiates network capabilities at inference time. The network performing sequential inference, referred to as the Classifier, consists of Long Short-Term Memory (LSTM) units and receives Z and $\mathbf{o}_{1:T}$. We train the network's goal belief B_{tg}^{Net} to converge to B_{tg}^{Joint} with loss:

$$\mathcal{L}_{\text{FR}} = \left(\frac{1}{2}\mathcal{D}_{\text{KL}}(B_{tg}^{\text{Joint}} \parallel B_{tg}^{\text{Net}}) + \frac{1}{2}\mathcal{D}_{\text{KL}}(B_{tg}^{\text{Net}} \parallel B_{tg}^{\text{Joint}})\right)^{\alpha},$$

where $\alpha = 0.5$ improves training efficiency. We use network output as marginal belief *updates* $\Delta B_t \in \mathbb{R}^{C\times R}$ (Figure 3a), where r_g can be estimated with a sum of log likelihood updates:

$$\arg\max_{r_g} P_Z(r_g \mid \mathbf{o}_{1:T}) = \arg\max_{r_g}\left(\sum_{\mathbf{r}_{\setminus g}} \exp\left(\sum_{t=1}^{T} \ln P_Z(\mathbf{o}_t \mid \mathbf{r})\right)\right).$$

Notice that the summation over non-goal realizations $\mathbf{r}_{\setminus g}$ envelops the temporal accumulation of a log probability. Since this cannot be decomposed into a summation over marginal likelihoods, this is Markovian only with respect to the *joint* posterior. By circumventing the challenge of integration, this setup isolates the capacity of memory dynamics to *re-evaluate* marginal beliefs with history-dependent corrections, which can be functionally equivalent to updating the joint posterior (A.3). In Figure 3b-d, the Fully Trained RNN reaches optimality, whereas the ESN, trained towards the same target, plateaus precisely at Naive Bayes. As predicted, when $C = 1$ and FR is zero, the gap between networks vanishes. The performance benefit of task-aligned dynamics (Schuessler et al., 2024) is apparent in the relative accuracy between networks, which is identical to the ratio between Bayesian observers. These results still persist when the number of trainable parameters is matched and even when we use ESN output to update joint beliefs (Supplementary Figure 9).

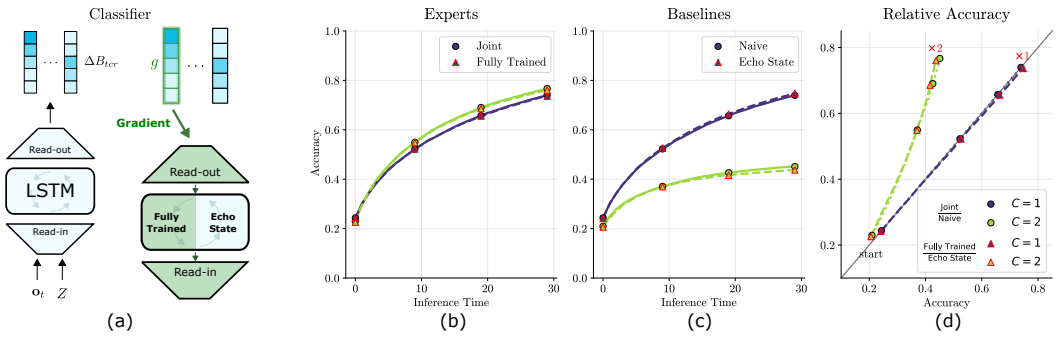

Figure 3: **RNNs align with theoretical predictions**. **(a)** Architecture (left) and gradient flow (right) of the Classifier. Only the goal belief receives a gradient. **(b-d)** Same as Figure 2b for $C = 1, 2$ ($C = 3$ in Supplementary Figure 10) with Fully Trained RNNs and ESNs (averaged over $n = 5$ random seeds). Markers indicate four equidistant reference time-points throughout inference.

Naive Bayes predicts the failure mode exhibited by Reservoir Computing. Next, we take a closer look at the belief dynamics in $C = 2$, where Naive Bayes predicts a unique failure mode analogous to hallucination. Figure 4a shows a representative trajectory of goal belief under Joint and Naive Bayes where $P(Hit) = 1-P(Miss)$ is the belief assigned to the correct action. In Figure 4b, we pool $P(Hit)$ over many episodes and analyze its density at each step t of inference. Early on, the distribution of *Hits* is centered near chance, but as time progresses (blue $\rightarrow$ red) the density shifts rightward, reflecting a growing $P(Hit)$. Interestingly, in the right column a second, disjoint peak below chance emerges. This reflects a significant portion of trajectories where misinterpretation makes evidence *harmful* to Naive Bayes and ESNs, an effect that grows with FR (Supplementary Figure 11). In B.3 we explore the hypothesis that Fully Trained RNNs stay reliable by adapting their dynamics to the history-dependence of the optimal marginal belief update.

4.3 Experiment 2: Compositional world structure can be learned indirectly

Thus far, we have established that neglecting Factorization Regret impedes optimal decision-making. However, networks in Experiment 1 were trained with explicit access to Z. Now we treat Z as unknown but learnable. Consider a large set $\mathcal{S}$ of latent variables, where the possible realizations of

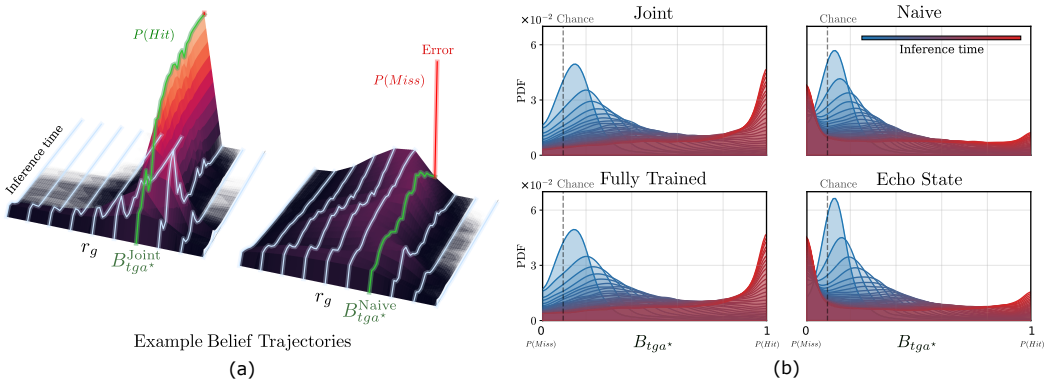

Figure 4: **Failure to capture FR can induce hallucinations**. **(a)** Sequential updating of marginal beliefs under Joint and Naive inference. **(b)** Distribution of belief assigned to the correct action at each step of inference, pooled over episodes. Colors progress from blue to red throughout inference.

variable $s \in \mathcal{S}$ represent a set of mutually exclusive realizations of the world, e.g. different seasons. Depending on the season, sunlight in the morning could mean that your alarm will trigger soon or that you've already slept through it. We assume that each variable carries a fixed set of vectors that encode how it interacts with other variables. Thus, learning how season interacts with time of day helps predict how season will interact with other variables. We call these vectors *Compositional Embeddings*. Formally, we associate variable s with a fixed set of embedding vectors $\{(\mathbf{k}_s^i, \mathbf{q}_s^i)\}_{i=1}^{d_o} \in \mathcal{E}$, and compute interactions $z_{cc'}^i = (\mathbf{k}_c^i)^\top \mathbf{q}_{c'}^i$ between all distinct pairs c, c' of relevant variables in the context $\mathcal{C} \subset \mathcal{S}$ (where $|\mathcal{C}| = C$). Since our focus is generalization to unseen combinations, we treat variable identities $c, c' \in \mathcal{C}$ as known and treat 10% of all variables as *testing variables*. Episodes are split into *training episodes*, which contain at most one testing variable (never the goal), and *testing episodes*, which consist entirely of *testing variables*. Since belief over a testing variable never receives a gradient, its embeddings must be learned indirectly (Figure 5).

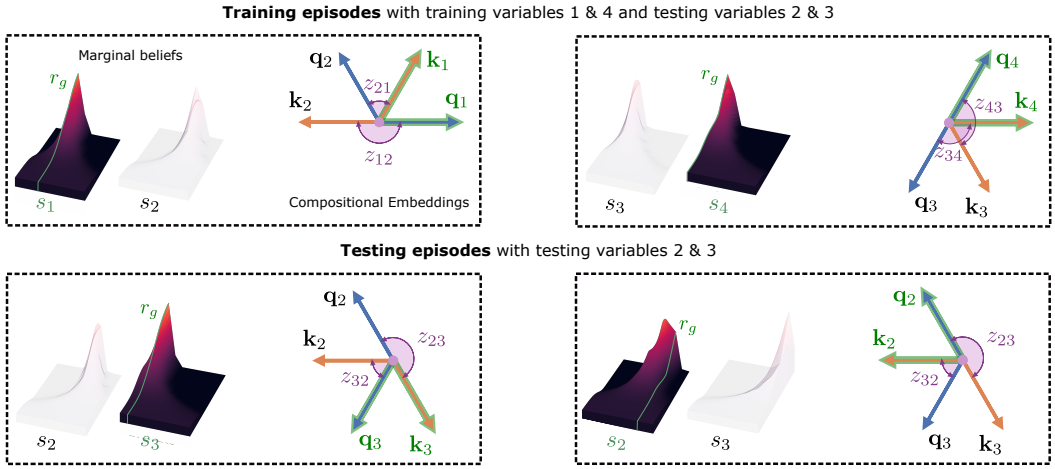

Figure 5: **Compositional generalization schematic.** Testing variable embeddings must be learned through their implicit relationships to training variables. Goal embeddings are outlined in green.

Naturalistic feedback is sufficient for learning interactions. By conditioning a generative model $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}})$ on estimated interactions $\hat{Z}$, a gradient can be propagated back to learnable embeddings $\{(\hat{\mathbf{k}}_s^i, \hat{\mathbf{q}}_s^i)\}_{i=1}^{d_o} \in \hat{\mathcal{E}}$ (Figure 6a). Estimated realizations $\hat{\mathbf{r}}$ are sampled from the Classifier's final belief B_T (A.4). For $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}})$ to be learnable, $\hat{\mathbf{r}}$ must be reliable. This presents a dilemma: learning embeddings requires accurate inference, yet accurate inference relies on the learned embeddings. Motivated by V-JEPA 2-AC (Assran et al., 2025), we separate the challenges of representation and

classification. *Representation Classification Chain* (RCC) (Figure 6b) modules bootstrap off each other without sharing gradients: a Classifier uses $\hat{Z}$ to infer $\hat{\mathbf{r}}$, while a Generator uses $\hat{\mathbf{r}}$ to reconstruct observations (A.4). The Classifier uses its goal belief as a policy, sampling one action at the end of each trajectory $a \sim B_{Tgr}$ to acquire sparse, partial, and binary feedback depending on whether or not the selected action was optimal $\delta(a - a^\star) \in \{0, 1\}$:

$$\mathcal{L}_{\text{Classifier}} \approx -\frac{1}{T}\sum_t \big(\delta \ln B_{tga} + (1-\delta)\ln(1 - B_{tga})\big).$$

Concurrently, the Generator minimizes a Jeffreys Divergence, $J(p \,\|\, q) = \mathcal{D}_{\text{KL}}(p \parallel q) + \mathcal{D}_{\text{KL}}(q \parallel p)$, between the empirical trajectory distribution $\tilde{P}(\mathbf{o})$ and its own predictions $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}})$:

$$\mathcal{L}_{\text{Generator}} = J\big(\tilde{P}(\mathbf{o}) \,\|\, \hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}})\big),$$

where the gradient flows from $\hat{Z}$ to $\hat{\mathcal{E}}$, teaching the embeddings to be useful for prediction. Figure 6c demonstrates that $\hat{\mathcal{E}}$ is indeed learnable from naturalistic feedback. Interestingly, this generalization only emerges once $|\mathcal{S}|$ is large enough. When the world contains too few latent variables, training variable embeddings are learned, but this does not transfer to testing variables (Supplementary Figure 12). Ablating the Generator and using gradient flow from the Classifier significantly impairs learning (Supplementary Figure 13).

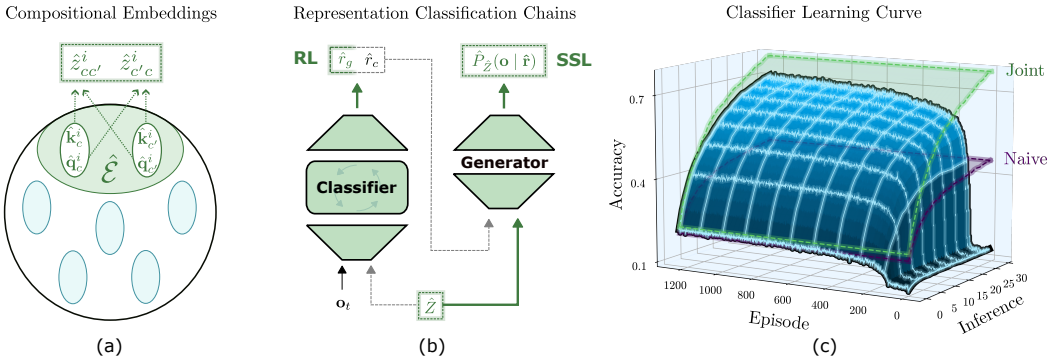

Figure 6: **A variational architecture learns Compositional Embeddings from reward. (a)** Relevant variables interact via learnable embeddings to form interactions. **(b)** Forward pass and gradient flow of the Classifier and Generator. The Classifier learns from rewards while the Generator uses self-supervised learning (SSL). **(c)** Testing episode accuracy of the Classifier throughout training.

4.4 Experiment 3: Latent space navigation enables zero-shot control

There is a developing interest in the generalization that emerges from conditioning on variables such as action, context, cluster, class, and constraints (Hu et al., 2018; Rakelly et al., 2019; Ajay et al., 2022; Hafner et al., 2025). In the previous section, we demonstrated that conditioning on interactions enables generalization to novel combinations of latent variables. Building on this, we show how learned Compositional Embeddings facilitate optimal behavior at the first opportunity for control, such as using a brand new oven. In this setup we consider an agent that has control over the latent variables $\mathbf{r}$, modeling a hypothetical interaction with the world. Note that the action space has expanded to the entire latent space. A *preferred observation* $\boldsymbol{\Omega}$ indicates success (such as the smell of a freshly baked pizza) and associates each location in latent space with *intrinsic reward* $\mathcal{R}$:

$$\mathcal{R}(\mathbf{r}, \boldsymbol{\Omega}, \ell) = \exp\left(\frac{1}{d_o}\sum_i \big(\ln(\ell_{\mathbf{z}^i}(\mathbf{r}))\Omega^i + \ln(1 - \ell_{\mathbf{z}^i}(\mathbf{r}))(1 - \Omega^i)\big)\right).$$

We can visualize the dependence of $\mathcal{R}(\mathbf{r}, \boldsymbol{\Omega}, \ell)$ on $\mathbf{r}$ as a *preference landscape* where the task is to find its maxima— the action most likely to generate a preferred observation (Figure 7, left). For a given $\boldsymbol{\Omega}$, a simple (and memory-less) controller can optimize a context-dependent policy $\pi(\mathbf{r} \mid \hat{Z})$ and value function $\nu(\hat{Z})$, with a standard Advantage Actor-Critic (A2C) loss:

$$\mathcal{L}_{A2C} \approx A^2 - A_{\text{Detached}} \ln \pi(\mathbf{r} \mid \hat{Z}) + \beta_{cntrl} \textstyle\sum_{\mathbf{r}'} \pi(\mathbf{r}' \mid \hat{Z}) \ln \pi(\mathbf{r}' \mid \hat{Z}),$$

where advantage $A = \mathcal{R}(\mathbf{r}, \boldsymbol{\Omega}, \ell) - \nu(\hat{Z})$ and β_{cntrl} is an entropy bonus (see subsection A.5).

Training in imagination. Notice that intrinsic reward $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$ depends on ℓ. The RCC can approximate ℓ with its learned $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \mathbf{r})$, which allows the controller to be trained offline in imagined environments. This inverts offline model-based optimization (Trabucco et al., 2021), which learns the objective function from a fixed set of evaluated observations. In our setting, the RCC imagines and evaluates observations to help a controller navigate latent space. Figure 7 (right) displays several example trajectories of $\arg\max_{\mathbf{r}} \pi(\mathbf{r} \mid \hat{Z})$ throughout learning, overlaid on the true preference landscape, showing that the controller identifies the optimal action without any interactions with the environment. Despite offline optimization's high sensitivity to model errors (Ha & Schmidhuber, 2018; Trabucco et al., 2021), offline training with ℓ and $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \mathbf{r})$ is indistinguishable (A.5).

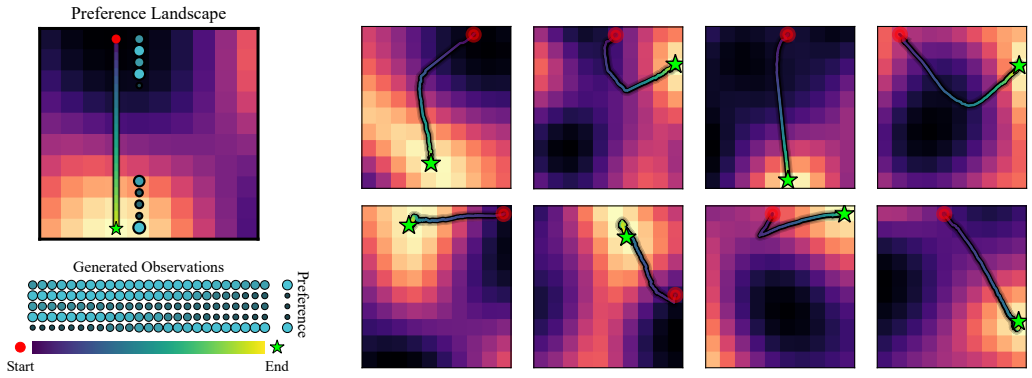

Figure 7: **Latent space navigation enables zero-shot control over observations.** Schematic illustration (left) and empirical trajectories (right) of an agent navigating a novel latent space to identify the state of the world which would generate the maximally preferred observation. Heatmaps indicate $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$ and trajectories show $\arg\max_{\mathbf{r}} \pi(\mathbf{r} \mid \hat{Z})$ throughout training (smoothed for clarity).

5 Discussion

Interactions push us to discover more than just our goals, yet their combinatorial growth renders the world intractable. We introduced Factorization Regret to measure the goal-relevant (*semantic*) information that is lost when interactions are ignored, and the Cognitive Gridworld as a setting that generates FR while remaining tractable for analytical treatment. FR explained the performance gap between Fully Trained RNNs and ESNs, and predicted a failure mode that decoupled confidence from accuracy. Because Naive Bayes is the best one can do before learning how variables interact, it offers a principled threshold for *meaningful* intelligence: crossing this threshold means an agent has used its internal representation to move beyond a stimulus-bound strategy (subsection B.2).

Since FR is a property of the environment, it could quantify how much of an existing benchmark's difficulty lies in learning interactions. In other compositional benchmarks, such as Alchemy (Wang et al., 2021) and CompoSuite (Mendez et al., 2022), task structure is explicit or discovered by manipulating the world, which leaves open how much their difficulty stems from challenges beyond compositionality. In the Cognitive Gridworld all learnable structure is latent, and it can be learned without causal manipulations. This allows us to quantify the task difficulty that stems from our intended source. We presented RCCs as an existence proof that the task's compositional latent structure can be learned from naturalistic feedback, and despite our emphasis on latent space, we also demonstrate that this setting is relevant to zero-shot control.

However, the scope of both the Cognitive Gridworld and RCCs is currently limited. The spatiotemporal richness of sensory observations was abstracted away entirely, and could be reintroduced through simple latent variable dynamics. Interactions between variables could be made richer by introducing sparsity to compositional embeddings, and our tensor likelihood construction extends naturally to multiple frequencies in a discrete Fourier series. RCCs were only evaluated in small ($C = 2$) environments and their scalability remains unknown. Future work should also implement additional sequence models, such as Transformers, which could vary in their mechanistic implementation of history-dependent belief updates (see subsection B.3 for a preliminary investigation

into the RNN implementation). The most pressing limitation is the assumption that all task-relevant variables are known, which a latent variable discovery algorithm (e.g. VASE) could solve.

Beyond artificial intelligence, this framework may also help explain how animals acquire reliable intuitions about the world. Assuming known variables and a stationary world was motivated by the hypotheses that the HPC discovers latent and temporal structure (Whittington et al., 2020; Raju et al., 2024). In this work, we focused on a complementary nuance: inference over interacting variables in a fixed, but unknown, environment. We propose that handling latent variable interactions is a major function of the Orbitofrontal Cortex (OFC), a region that is highly specialized in primates (Preuss & Wise, 2022). The OFC is involved in latent state inference, context-dependent behavior and compositionality (Duarte et al., 2010; Takahashi et al., 2013; Wilson et al., 2014; Farovik et al., 2015; Schuck et al., 2016; Chan et al., 2016; Pelletier & Fellows, 2019; Stalnaker et al., 2021; Mizrak et al., 2021; Schiereck et al., 2024). Its activity is thought to encode a "common currency" value, including the value of information and imagined outcomes, the location of spatial goals, and a multidimensional abstract space with reward-dependent compression (Padoa-Schioppa & Assad, 2008; Padoa-Schioppa, 2009; Kahnt et al., 2010; Takahashi et al., 2013; Stalnaker et al., 2015; Basu et al., 2021; Muhle-Karbe et al., 2023; Qiu et al., 2024; Bussell et al., 2024). Damage to the region has long been associated with impaired cognitive flexibility and re-evaluation (Butter, 1969; Dias et al., 1996; Schoenbaum et al., 2002; McAlonan & Brown, 2003; Chudasama & Robbins, 2003; Hornak et al., 2004; Izquierdo et al., 2004; Ryguła et al., 2010; Rudebeck et al., 2013; Murray et al., 2015), particularly when value must be inferred, but not when cached values suffice (Jones et al., 2012). These seemingly disparate roles in inference and evaluation would be unified by the objective of managing Factorization Regret.

The HPC and OFC are thought to maintain complementary and interacting cognitive maps (Wikenheiser & Schoenbaum, 2016; Wikenheiser et al., 2017; Knudsen & Wallis, 2020; Wang et al., 2020; Mizrak et al., 2021; Lin & Zhou, 2024). The HPC is well suited to dynamic environments in which the order of events matters (Raju et al., 2024). We suggest that the HPC compresses ordered sensory sequences into Compositional Embeddings that are consolidated throughout the cortex. Yet once the spatiotemporal richness of the sensory world has been abstracted away, the task of belief updating still remains. Given a novel goal, the brain could recruit distributed cortical representations into a cognitive map sufficient for approximate inference of the optimal action, i.e., meta-learning. We propose that the OFC orchestrates this process, using its robust connections with limbic, thalamic, and neuromodulatory systems to influence inter-cortical communication (Rempel-Clower, 2007; Gao et al., 2007; Takahashi et al., 2009; Lodge, 2011; Jo & Mizumori, 2016; Namboodiri et al., 2019; Weitz et al., 2019) and evaluate imagined outcomes (as in subsection 4.4). Furthermore, by detecting unresolved Factorization Regret, the OFC could identify when new interactions should be learned by the HPC. We therefore hypothesize that detecting and managing Factorization Regret is a collaborative objective of the OFC and HPC: the HPC consolidates latent variable embeddings throughout the cortex, while the OFC recruits these representations and expands their interactions into a brain-wide cognitive map for zero-shot control.

The barrier to generalization once all relevant variables are known is understanding how to put them together. The presence of Factorization Regret indicates that learning to infer the latent state of the world could be helpful for more than just the immediate goal, and therefore environments with greater FR exert greater pressure to develop compositional representations. Rather than describing intelligence as a computation to solve a problem, this perspective casts intelligence as a property selected for by the environment. This then begs the question: what kinds of environments selected for *meaningful* intelligence in animals, and can we build them?

Reproducibility Statement

Detailed environment and model hyperparameters, extended generative process descriptions, network architectures, and pseudo-code are provided in the Appendices. The full code used to generate the results of this paper will be made available in an open source repository.

Ethics Statement

This work involved no human subjects, personal data, or deployed systems.

REFERENCES

Alessandro Achille, Tom Eccles, Loïc Matthey, Christopher P. Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:52049801.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, T. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv*, abs/2211.15657, 2022. URL https://api.semanticscholar.org/CorpusID:254044710.

Sergio A Alvarez. An exact analytical relation among recall, precision, and classification accuracy in information retrieval. Technical Report BC-CS-2002-01, Computer Science Department, Boston College, 2002.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv*, abs/2506.09985, 2025. URL https://api.semanticscholar.org/CorpusID:279306055.

Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. URL https://api.semanticscholar.org/CorpusID:121222106.

Omri Barak, Mattia Rigotti, and Stefano Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *The Journal of Neuroscience*, 33:3844 – 3856, 2013. URL https://api.semanticscholar.org/CorpusID:1766932.

Raunak Basu, Robert Gebauer, Tim Herfurth, Simon Kolb, Zahra Golipour, Tatjana Tchumatchenko, and Hiroshi T. Ito. The orbitofrontal cortex maps future navigational goals. *Nature*, 599:449 – 452, 2021. URL https://api.semanticscholar.org/CorpusID:240072183.

Timothy Edward John Behrens, Timothy H. Muller, James C. R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100:490–509, 2018. URL https://api.semanticscholar.org/CorpusID:53105626.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

Gauthier Boeshertz and Claudia Clopath. Predictive learning enables compositional representations. *bioRxiv*, 2025.

John C. Bowler, Dua Azhar, Cambria M Jensen, Hyun-Woo Lee, and James G. Heys. Structured experience shapes strategy learning and neural dynamics in the medial entorhinal cortex. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:278664552.

Bingni W. Brunton, Matthew M. Botvinick, and Carlos D. Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340:95 – 98, 2013. URL https://api.semanticscholar.org/CorpusID:13098239.

Yoram Burak. Spatial coding and attractor dynamics of grid cells in the entorhinal cortex. *Current Opinion in Neurobiology*, 25:169–175, 2014. URL https://api.semanticscholar.org/CorpusID:16681043.

Jennifer J. Bussell, Ryan P. Badman, Christian David Márton, Ethan S. Bromberg-Martin, Larry Abbott, Kanaka Rajan, and Richard Axel. Representations of the intrinsic value of information in mouse orbitofrontal cortex. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:264171514.

Charles M. Butter. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in Macaca mulatta. *Physiology & Behavior*, 4:163–171, 1969. URL https://api.semanticscholar.org/CorpusID:17920166.

Stephanie C. Y. Chan, Yael Niv, and Kenneth A. Norman. A probability distribution over latent causes, in the orbitofrontal cortex. *The Journal of Neuroscience*, 36:7817 – 7828, 2016. URL https://api.semanticscholar.org/CorpusID:9673546.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34:15084–15097, 2021.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. ARC Prize 2024: Technical report. *arXiv*, abs/2412.04604, 2024. URL https://api.semanticscholar.org/CorpusID:274581906.

Yogita Chudasama and Trevor William Robbins. Dissociable contributions of the orbitofrontal and infralimbic cortex to Pavlovian autoshaping and discrimination reversal learning: Further evidence for the functional heterogeneity of the rodent frontal cortex. *The Journal of Neuroscience*, 23:8771 – 8780, 2003. URL https://api.semanticscholar.org/CorpusID:8018743.

Mark M. Churchland and Krishna V. Shenoy. Preparatory activity and the expansive nullspace. *Nature Reviews Neuroscience*, 2024. URL https://api.semanticscholar.org/CorpusID:268250917.

Joni Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2, 2012. URL https://api.semanticscholar.org/CorpusID:7342429.

Victor de Lafuente, Mehrdad Jazayeri, and Michael N. Shadlen. Representation of accumulating evidence for a decision in two parietal areas. *The Journal of Neuroscience*, 35:4306 – 4318, 2015. URL https://api.semanticscholar.org/CorpusID:14214715.

Rebecca Dias, Trevor William Robbins, and Angela C. Roberts. Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, 380:69–72, 1996. URL https://api.semanticscholar.org/CorpusID:4301013.

Audrey Duarte, Richard N. A. Henson, Robert T. Knight, Tina Emery, and Kim S. Graham. Orbito-frontal cortex is necessary for temporal context memory. *Journal of Cognitive Neuroscience*, 22:1819–1831, 2010. URL https://api.semanticscholar.org/CorpusID:14909943.

Alexis M. Dubreuil, Adrian Valente, Manuel Beiran, Francesca Mastrogiuseppe, and Srdjan Ostojic. The role of population structure in computations through neural dynamics. *Nature Neuroscience*, 25:783 – 794, 2022. URL https://api.semanticscholar.org/CorpusID:256838997.

Anja Farovik, Ryan Place, Sam McKenzie, Blake S. Porter, Catherine E Munro, and Howard Eichenbaum. Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. *The Journal of Neuroscience*, 35:8333 – 8344, 2015. URL https://api.semanticscholar.org/CorpusID:17512263.

Ming Gao, Chang-Liang Liu, Shen Yang, Guo-Zhang Jin, Benjamin S Bunney, and Wei-Xing Shi. Functional coupling between the prefrontal cortex and dopamine neurons in the ventral tegmental area. *Journal of Neuroscience*, 27(20):5414–5421, 2007.

Mona M. Garvert, Tankred Saanum, Eric Schulz, Nicolas W. Schuck, and Christian F. Doeller. Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nature Neuroscience*, 26:615 – 626, 2023. URL https://api.semanticscholar.org/CorpusID:257924320.

Samuel J. Gershman and Yael Niv. Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, 20:251–256, 2010. URL https://api.semanticscholar.org/CorpusID:10255984.

AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1326–1330, 2017. URL https://api.semanticscholar.org/CorpusID:8283977.

Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–74, 2007. URL https://api.semanticscholar.org/CorpusID:6842034.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647 – 653, 2025. URL https://api.semanticscholar.org/CorpusID:277508993.

Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 12, 2018. URL https://api.semanticscholar.org/CorpusID:57761278.

Steven C Hayes, Dermot Barnes-Holmes, and Bryan Roche. *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science & Business Media, 2001.

Jay A. Hennig, Sandra A. Romero Pinto, Takahiro Yamaguchi, Scott W. Linderman, Naoshige Uchida, and Samuel J. Gershman. Emergence of belief-like representations through reinforcement learning. *PLOS Computational Biology*, 19, 2023. URL https://api.semanticscholar.org/CorpusID:258051351.

Jeffrey R. Hollerman and Wolfram Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309, 1998. URL https://api.semanticscholar.org/CorpusID:7785929.

J. Hornak, John P. O'Doherty, Jessica Bramham, Edmund T. Rolls, Robin G. Morris, Peter R. Bullock, and C. E. Polkey. Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience*, 16:463–478, 2004. URL https://api.semanticscholar.org/CorpusID:132678.

Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169, 2022. URL https://api.semanticscholar.org/CorpusID:215744839.

Peiyun Hu, Zachary Chase Lipton, Anima Anandkumar, and Deva Ramanan. Active learning with partial feedback. *arXiv*, abs/1802.07427, 2018. URL https://api.semanticscholar.org/CorpusID:3534906.

Masanobu Inubushi and Kazuyuki Yoshimura. Reservoir computing beyond memory-nonlinearity trade-off. *Scientific Reports*, 7, 2017. URL https://api.semanticscholar.org/CorpusID:10886282.

Alicia Izquierdo, Robin K. Suda, and Elisabeth A. Murray. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *The Journal of Neuroscience*, 24:7540 – 7548, 2004. URL https://api.semanticscholar.org/CorpusID:17542448.

Yong Sang Jo and Sheri J. Y. Mizumori. Prefrontal regulation of neuronal activity in the ventral tegmental area. *Cerebral Cortex*, 26(10):4057–4068, 2016. URL https://api.semanticscholar.org/CorpusID:4875389.

Joshua L. Jones, Guillem R. Esber, Michael A. McDannald, Aaron J. Gruber, Alex Hernandez, Aaron Mirenzi, and Geoffrey Schoenbaum. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, 338(6109):953–956, 2012.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101:99–134, 1998. URL https://api.semanticscholar.org/CorpusID:5613003.

Thorsten Kahnt, Jakob Heinzle, Soyoung Q. Park, and John-Dylan Haynes. The neural code of reward anticipation in human orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, 107:6010 – 6015, 2010. URL https://api.semanticscholar.org/CorpusID:22879670.

Yul HR Kang, Frederike H. Petzschner, Daniel M. Wolpert, and Michael N. Shadlen. Piercing of consciousness as a threshold-crossing operation. *Current Biology*, 27:2285 – 2295.e6, 2017. URL https://api.semanticscholar.org/CorpusID:27618011.

Gili Karni, Yael Niv, and Nathaniel Daw. How goals affect information seeking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

Kenneth Kay, Natalie Biderman, Ramin Khajeh, Manuel Beiran, Christopher J. Cueva, Daphna Shohamy, Greg Jensen, Xue-Xin Wei, Vincent P. Ferrera, and L.F. Abbott. Emergent neural dynamics and geometry for generalization in a transitive inference task. *PLOS Computational Biology*, 20(4):e1011954, 2024. URL https://api.semanticscholar.org/CorpusID:260381252.

Roozbeh Kiani and Michael N. Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324:759 – 764, 2009. URL https://api.semanticscholar.org/CorpusID:11581812.

Eric B. Knudsen and Joni D. Wallis. Closed-loop theta stimulation in the orbitofrontal cortex prevents reward-based learning. *Neuron*, 2020. URL https://api.semanticscholar.org/CorpusID:212644121.

Artemy Kolchinsky and David H. Wolpert. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8, 2018. URL https://api.semanticscholar.org/CorpusID:53566383.

Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623:115 – 121, 2023. URL https://api.semanticscholar.org/CorpusID:264489248.

Andrew Kyle Lampinen, Martin Engelcke, Yuxuan Li, Arslan Chaudhry, and James L. McClelland. Latent learning: episodic memory complements parametric learning by enabling flexible reuse of experiences. *arXiv preprint arXiv:2509.16189*, 2025. URL https://api.semanticscholar.org/CorpusID:281410976.

Denis C L Lan, Laurence T Hunt, and Christopher Summerfield. Goal-directed navigation in humans and deep reinforcement learning agents relies on an adaptive mix of vector-based and transition-based strategies. *PLOS Biology*, 23, 2025. URL https://api.semanticscholar.org/CorpusID:280389881.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Huixin Lin and Jingfeng Zhou. Hippocampal and orbitofrontal neurons contribute to complementary aspects of associative structure. *Nature Communications*, 15, 2024. URL https://api.semanticscholar.org/CorpusID:270638438.

Daniel J. Lodge. The medial prefrontal and orbitofrontal cortices differentially regulate dopamine system function. *Neuropsychopharmacology*, 36:1227–1236, 2011. URL https://api.semanticscholar.org/CorpusID:28747941.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503:78 – 84, 2013. URL https://api.semanticscholar.org/CorpusID:4450696.

Kerry McAlonan and Verity J. Brown. Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. *Behavioural Brain Research*, 146:97–103, 2003. URL https://api.semanticscholar.org/CorpusID:11359123.

Jorge A Mendez, Marcel Hussing, Meghna Gummadi, and Eric Eaton. CompoSuite: A compositional reinforcement learning benchmark. *arXiv preprint arXiv:2207.04136*, 2022.

Kevin Miller, Maria Eckstein, Matt Botvinick, and Zeb Kurth-Nelson. Cognitive model discovery via disentangled RNNs. *Advances in Neural Information Processing Systems*, 36:61377–61394, 2023.

Eda Mizrak, Nichole R. Bouffard, Laura A. Libby, Erie D. Boorman, and Charan Ranganath. The hippocampus and orbitofrontal cortex jointly represent task structure during memory-guided decision making. *Cell Reports*, 37:110065 – 110065, 2021. URL https://api.semanticscholar.org/CorpusID:244792239.

George E. Monahan. State of the art—A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28:1–16, 1982. URL https://api.semanticscholar.org/CorpusID:123582406.

Paul S. Muhle-Karbe, Hannah Sheahan, Giovanni Pezzulo, Hugo J. Spiers, Samson Chien, Nicolas W. Schuck, and Christopher Summerfield. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, 111:3885–3899.e6, 2023. URL https://api.semanticscholar.org/CorpusID:255850293.

Elisabeth A. Murray, Emily J. Moylan, Kadharbatcha S. Saleem, Benjamin M. Basile, and Janita Turchi. Specialized areas for value updating and goal selection in the primate orbitofrontal cortex. *eLife*, 4:e11695, 2015.

Vijay Mohan K. Namboodiri, James M. Otis, Kay van Heeswijk, Elisa S Voets, Rizk A. Alghorazi, Jose Rodríguez-Romaguera, Stefan Mihalas, and Garret D. Stuber. Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nature Neuroscience*, 22:1110 – 1121, 2019. URL https://api.semanticscholar.org/CorpusID:173992358.

Yael Niv, Reka Daniel, Andra Geana, Samuel J. Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C. Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, 35:8145 – 8157, 2015. URL https://api.semanticscholar.org/CorpusID:18446484.

Pedro A. Ortega, Jane X. Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Manfred Otto Heess, Joel Veness, Alexander Pritzel, Pablo Sprechmann, Siddhant M. Jayakumar, Tom McGrath, Kevin J. Miller, Mohammad Gheshlaghi Azar, Ian Osband, Neil C. Rabinowitz, András György, Silvia Chiappa, Simon Osindero, Yee Whye Teh, Hado van Hasselt, Nando de Freitas, Matthew M. Botvinick, and Shane Legg. Meta-learning of sequential strategies. *arXiv*, abs/1905.03030, 2019. URL https://api.semanticscholar.org/CorpusID:147703875.

Camillo Padoa-Schioppa. Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of Neuroscience*, 29:14004 – 14014, 2009. URL https://api.semanticscholar.org/CorpusID:7643973.

Camillo Padoa-Schioppa and John A Assad. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature Neuroscience*, 11:95–102, 2008. URL https://api.semanticscholar.org/CorpusID:901185.

Gabriel Pelletier and Lesley K. Fellows. A critical role for human ventromedial frontal lobe in value comparison of complex objects based on attribute configuration. *The Journal of Neuroscience*, 39:4124 – 4132, 2019. URL https://api.semanticscholar.org/CorpusID:76659569.

Michael Pereira, Pierre Mégevand, Mi Xue Tan, Wenwen Chang, Shuo Wang, Ali Rezai, Margitta Seeck, Marco Vincenzo Corniola, Shahan Momjian, Fosco Bernasconi, Olaf Blanke, and Nathan Faivre. Evidence accumulation relates to perceptual consciousness and monitoring. *Nature Communications*, 12, 2021. URL https://api.semanticscholar.org/CorpusID:235268827.

Philip and Hemang. SimpleBench: The Text Benchmark in which Unspecialized Human Performance Exceeds that of Current Frontier Models. https://simple-bench.com/, October 2024. Technical Report.

Todd M. Preuss and Steven P. Wise. Evolution of prefrontal cortex. *Neuropsychopharmacology*, 47(1):3–19, 2022. URL https://api.semanticscholar.org/CorpusID:236940889.

Alexandra Proca, Fernando E. Rosas, Andrea I. Luppi, Daniel Bor, Matthew Crosby, and Pedro A. M. Mediano. Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *PLOS Computational Biology*, 20, 2024. URL https://api.semanticscholar.org/CorpusID:252734834.

Yidan Qiu, Huakang Li, Jiajun Liao, Kemeng Chen, Xiaoyan Wu, Bingyi Liu, and Ruiwang Huang. Forming cognitive maps for abstract spaces: the roles of the human hippocampus and orbitofrontal cortex. *Communications Biology*, 7, 2024. URL https://api.semanticscholar.org/CorpusID:269499660.

Rajkumar Vasudeva Raju, J. Swaroop Guntupalli, Guangyao Zhou, Carter Wendelken, Miguel Lázaro-Gredilla, and Dileep George. Space is a latent sequence: A theory of the hippocampus. *Science Advances*, 10, 2024. URL https://api.semanticscholar.org/CorpusID:271597277.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:84187276.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley D. Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *arXiv*, abs/2205.06175, 2022. URL https://api.semanticscholar.org/CorpusID:248722148.

Nancy Rempel-Clower. Role of orbitofrontal cortex connections in emotion. *Annals of the New York Academy of Sciences*, 1121, 2007. URL https://api.semanticscholar.org/CorpusID:21317263.

Fernando E. Rosas, Pedro A. M. Mediano, Michael Gastpar, and Henrik Jeldtoft Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100:032305, 2019. URL https://api.semanticscholar.org/CorpusID:67855406.

Nicholas Roy, Geoffrey J. Gordon, and Sebastian Thrun. Finding approximate POMDP solutions through belief compression. *Journal of Artificial Intelligence Research*, 23:1–40, 2005. URL https://api.semanticscholar.org/CorpusID:3894022.

Peter H. Rudebeck, Richard C. Saunders, Anna T. Prescott, Lily S. Chau, and Elisabeth A. Murray. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature Neuroscience*, 16(8):1140–1145, 2013.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2003. URL https://api.semanticscholar.org/CorpusID:262339885.

Rafał Ryguła, Susannah C Walker, Hannah F. Clarke, Trevor William Robbins, and Angela C. Roberts. Differential contributions of the primate ventrolateral prefrontal and orbitofrontal cortex to serial reversal learning. *The Journal of Neuroscience*, 30:14552 – 14559, 2010. URL https://api.semanticscholar.org/CorpusID:6290998.

Shervin Safavi and Peter Dayan. Multistability, perceptual value, and internal foraging. *Neuron*, 110:3076–3090, 2022. URL https://api.semanticscholar.org/CorpusID:251887280.

Shervin Safavi and Peter Dayan. A decision-theoretic model of multistability: perceptual switches as internal actions. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:274656763.

Adam Santoro, Sergey Bartunov, Matthew M. Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, 2016. URL https://api.semanticscholar.org/CorpusID:6466088.

Shannon S. Schiereck, Danilo Trinidad Pérez-Rivera, Andrew Mah, Margaret L. DeMaegd, Royall McMahon Ward, David Hocker, Cristina Savin, and Christine M. Constantinople. Neural dynamics in the orbitofrontal cortex reveal cognitive strategies. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:273776371.

Geoffrey Schoenbaum, Summer L. Nugent, Michael P. Saddoris, and B Setlow. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport*, 13:885–890, 2002. URL https://api.semanticscholar.org/CorpusID:23885025.

Nicolas W. Schuck, M. Cai, Robert C. Wilson, and Yael Niv. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91:1402–1412, 2016. URL https://api.semanticscholar.org/CorpusID:7912639.

Friedrich Schuessler, Francesca Mastrogiuseppe, Srdjan Ostojic, and Omri Barak. Aligned and oblique dynamics in recurrent neural networks. *eLife*, 13:RP93060, 2024. URL https://api.semanticscholar.org/CorpusID:259937586.

Michael N. Shadlen and William T. Newsome. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86(4):1916–36, 2001. URL https://api.semanticscholar.org/CorpusID:272332.

B. F. Skinner. *Science and Human Behavior*. Macmillan, New York, 1953. URL https://api.semanticscholar.org/CorpusID:270200242.

Thomas A. Stalnaker, Nisha K. Cooch, and Geoffrey Schoenbaum. What the orbitofrontal cortex does not do. *Nature Neuroscience*, 18(5):620–627, 2015.

Thomas A. Stalnaker, Nishika Raheja, and Geoffrey Schoenbaum. Orbitofrontal state representations are related to choice adaptations and reward predictions. *The Journal of Neuroscience*, 41:1941 – 1951, 2021. URL https://api.semanticscholar.org/CorpusID:231611820.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Yuji K. Takahashi, Matthew R. Roesch, Thomas A. Stalnaker, Richard Z. Haney, Donna J. Calu, Adam R. Taylor, Kathryn A. Burke, and Geoffrey Schoenbaum. The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron*, 62:269–280, 2009. URL https://api.semanticscholar.org/CorpusID:14014931.

Yuji K. Takahashi, Chun Yun Chang, Federica Lucantonio, Richard Z. Haney, Benjamin A. Berg, Hau-Jie Yau, Antonello Bonci, and Geoffrey Schoenbaum. Neural estimates of imagined outcomes in the orbitofrontal cortex drive behavior and learning. *Neuron*, 80:507–518, 2013. URL https://api.semanticscholar.org/CorpusID:1686176.

Joshua B. Tenenbaum. Building machines that learn and think like people. In *Adaptive Agents and Multi-Agent Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:260496023.

Peter Tiňo. Dynamical systems as temporal feature spaces. *J. Mach. Learn. Res.*, 21(44):1–42, 2020. URL https://api.semanticscholar.org/CorpusID:196621796.

Peter Tiňo, Robert Simon Fong, and Roberto F. Leonarduzzi. Predictive modeling in the reservoir kernel motif space. *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. URL https://api.semanticscholar.org/CorpusID:269756966.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv*, physics/0004057, 2000. URL https://api.semanticscholar.org/CorpusID:8936496.

Edward Chace Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. URL https://api.semanticscholar.org/CorpusID:42496633.

Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10358–10368. PMLR, 2021.

Elia Turner and Omri Barak. The simplicity bias in multi-task RNNs: Shared attractors, reuse of dynamics, and geometric representation. In *Neural Information Processing Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:268042264.

Christopher Versteeg, Jonathan D. McCart, Mitchell Ostrow, David M. Zoltowski, Clayton B. Washington, Laura Driscoll, Olivier Codol, Jonathan A. Michaels, Scott W. Linderman, David Sussillo, and Chethan Pandarinath. Computation-through-dynamics benchmark: Simulated datasets and quality metrics for dynamical models of neural activity. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:276259763.

Pietro Vertechi, Eran Lottem, Dario Sarra, Beatriz S. Godinho, Isaac Treves, Tiago Quendera, Matthijs N. Oude Lohuis, and Zachary F. Mainen. Inference-based decisions in a hidden state foraging task: Differential contributions of prefrontal cortical areas. *Neuron*, 106(1):166 – 176.e6, 2020. URL https://api.semanticscholar.org/CorpusID:211075310.

Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020. URL https://api.semanticscholar.org/CorpusID:216119345.

Fang Wang, Geoffrey Schoenbaum, and Thorsten Kahnt. Interactions between human orbitofrontal cortex and hippocampus support model-based inference. *PLoS Biology*, 18, 2020. URL https://api.semanticscholar.org/CorpusID:210862344.

Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew M. Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21:860 – 868, 2018. URL https://api.semanticscholar.org/CorpusID:44137923.

Jane X Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charlie Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, et al. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents. *arXiv preprint arXiv:2102.02926*, 2021.

Michael Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4:66–82, 1960. URL https://api.semanticscholar.org/CorpusID:9028505.

Andrew J. Weitz, Hyun Joo Lee, ManKin Choy, and Jin Hyung Lee. Thalamic input to orbitofrontal cortex drives brain-wide, frequency-dependent inhibition mediated by GABA and zona incerta. *Neuron*, 104:1153–1167.e4, 2019. URL https://api.semanticscholar.org/CorpusID:204837757.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy Edward John Behrens. The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183 (5):1249 – 1263.e23, 2020. URL https://api.semanticscholar.org/CorpusID:203374638.

Andrew M. Wikenheiser and Geoffrey Schoenbaum. Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, 17:513–523, 2016. URL https://api.semanticscholar.org/CorpusID:7047742.

Andrew M. Wikenheiser, Yasmin Marrero-Garcia, and Geoffrey Schoenbaum. Suppression of ventral hippocampal output impairs integrated orbitofrontal encoding of task structure. *Neuron*, 95:1197–1207.e3, 2017. URL https://api.semanticscholar.org/CorpusID:5342811.

Robert C. Wilson, Yuji K. Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81:267–279, 2014. URL https://api.semanticscholar.org/CorpusID:2713112.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22:297 – 306, 2019. URL https://api.semanticscholar.org/CorpusID:58006968.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

A METHODOLOGY

A.1 ENVIRONMENT HYPERPARAMETERS

- T (Trajectory / inference steps): 30
- $d_{\mathcal{E}}$ (Embedding dimensionality): 30
- $|\mathcal{S}|$ (Total latent variables): 500
- R (Possible realizations): 10
- d_o (Observation dimensions): 5
- λ (Likelihood temperature): 2

A.2 GENERATIVE PROCESS DETAILS

Compositional Embeddings are acquired by sampling $\mathbf{k}_s^i \sim \mathcal{N}(0,1)^{d_{\mathcal{E}}}, \quad \mathbf{q}_s^i \sim \mathcal{N}(0,1)^{d_{\mathcal{E}}}, \quad \forall s \in \mathcal{S}$, orthogonalizing across $\mathcal{O}$ via a Gram–Schmidt process and finally normalizing to 1 along $d_{\mathcal{E}}$. In each episode E, the generative process is described by:

$$o^i \sim P(o^i = 1 \mid \mathbf{k}_1^i, \mathbf{q}_1^i, \ldots, \mathbf{k}_C^i, \mathbf{q}_C^i, r_1, \ldots, r_C), \quad \forall i \in \mathcal{O},$$

which can be shortened to:

$$\ell_{\mathbf{z}^i}(\mathbf{r}) = \text{generation}(\underbrace{\underbrace{\text{compression}(\mathbf{k}_1^i, \mathbf{q}_1^i, \ldots, \mathbf{k}_C^i, \mathbf{q}_C^i)}_{\text{expansion}(z_1^i, \ldots, z_{C\times(C-1)}^i)}}_{\ell_{\mathbf{z}^i}}, r_1, \ldots, r_C)$$

Generalization to an *out of distribution* set $\{\mathbf{k}_1^i, \mathbf{q}_1^i, \ldots, \mathbf{k}_C^i, \mathbf{q}_C^i\}$ may occur if the resultant $\mathbf{z}^i$ falls within the manifold of training $\mathbf{z}^i$. The low rank structure of $\mathbf{z}^i \in \mathbb{R}^{C\times(C-1)}$ is embedded in joint likelihood, $\ell_{\mathbf{z}^i} \in [0,1]^{R^C}$. Below we step through the default expansion from $\mathbf{z}^i \to \ell_{\mathbf{z}^i}$ in 2 steps.

STEP 1. EXPANDING $z_{cc'}^i \to \mathbf{v}_{cc'}^i$:

First, we expand each scalar $z_{cc'}^i$ to vector $\mathbf{v}_{cc'}^i$ through Θ and ω:

$$\mathbf{v}_{cc'}^i(r) = \sum_{n=0}^{L-1} \Theta(n,r)\, \omega(n), \quad L = 1 + 2R.$$

Here, Θ is a cyclic shift of sinusoidal base values (indexing realizations from $r = 0$, a relabeling of the main text's $\{1, \ldots, R\}$):

$$\Theta(n,r) = \lambda \sin\!\left(\frac{2\pi}{L}\big(((n-r) \bmod L) + 1 - \frac{L}{2}\big)\right), \quad \lambda = \text{likelihood temperature},$$

and ω gives softmax weights over circular distances from $Lz_{cc'}^i$:

$$\omega(n) = \text{softmax}_n\big(-d(n)^2\big), \qquad d(n) = \min\big(m(n),\, L - m(n)\big),$$
$$\text{where} \quad m(n) = \big(n - Lz_{cc'}^i\big) \bmod L.$$

Functionally, $z_{cc'}^i$ performs a smooth approximation of the discrete *roll* operation of a sinusoid. Note that since $L > R$, $\mathbf{v}_{cc'}^i$ is shorter than a full period, allowing for asymmetry in $\mathbf{v}_{cc'}^i$.

STEP 2. EXPANDING $\mathbf{v}_{cc'}^i(\mathbf{v}_{c'c}^i)^\top \to \ell_{\mathbf{z}^i} \in [0,1]^{R^C}$:

Next, we expand outer products $\mathbf{v}_{cc'}^i(\mathbf{v}_{c'c}^i)^\top$ to tensor $\ell_{\mathbf{z}^i} \in [0,1]^{R^C}$ through:

$$\ell_{\mathbf{z}^i} \;=\; \begin{cases} \sigma\big(\mathbf{v}_{cc}^i\big), & C = 1, \\ \sigma\Big(\sum_{c<c'} \mathcal{B}_{cc'}\Big(\mathbf{v}_{cc'}^i(\mathbf{v}_{c'c}^i)^\top\Big)\Big), & C \geq 2, \end{cases}$$

where σ denotes element-wise sigmoid and $\mathcal{B}_{cc'} : \mathbb{R}^{R\times R} \to \mathbb{R}^{R^C}$ is a broadcasting operator that expands $\mathbb{R}^{R\times R}$ to $\mathbb{R}^{R^C}$ by adding singleton dimensions along all axes but c, c' of $\ell_{\mathbf{z}^i}$. For example, for the case of $C = 3$:

$$\ell_{\mathbf{z}^i} = \sigma\big(\{\mathbf{v}^i_{12}(\mathbf{v}^i_{21})^\top\}_{12} + \{\mathbf{v}^i_{13}(\mathbf{v}^i_{31})^\top\}_{13} + \{\mathbf{v}^i_{23}(\mathbf{v}^i_{32})^\top\}_{23}\big).$$

Materializing $\ell_{\mathbf{z}^i}$ for all i therefore requires $O(d_o R^C)$ memory, and each exact joint Bayesian update requires $O(d_o R^C)$ operations. For the network experiments ($R = 10$, $C \leq 3$) this is at most 5×10^3 entries (5×10^4 for the exact Bayesian analysis at $C = 4$ in Figure 2), but the exponential growth in C is what renders the joint posterior intractable in general and motivates approximate, marginal belief states. By contrast, the Classifier receives $d_o C(C-1)$ interactions (for $C \geq 2$) and outputs $C \times R$ marginal logits, and the Generator's inputs grow at most quadratically in C, so both scale polynomially; the exponential cost is confined to the likelihood tensor, exact joint inference, and the Controller's joint policy.

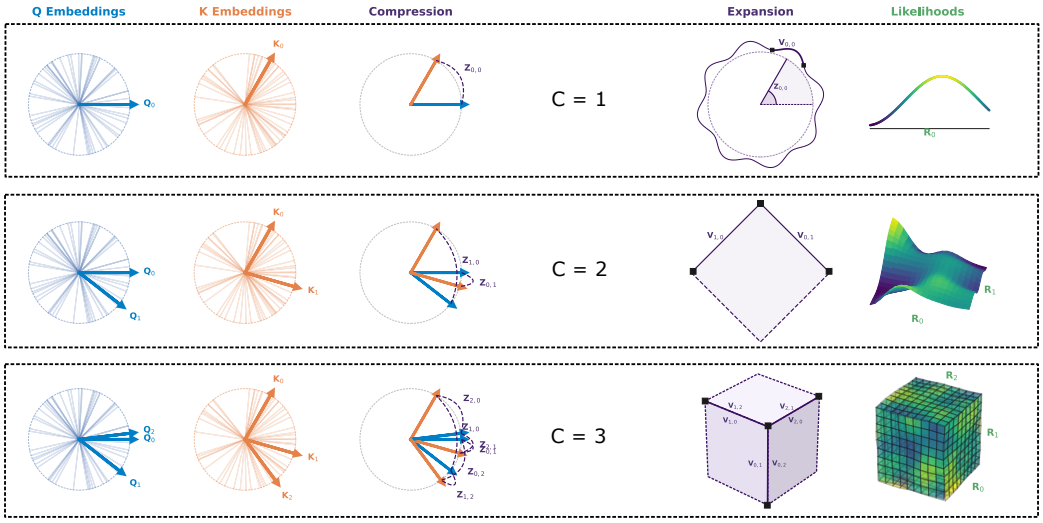

Figure 8: **A flexible process for embedding Gridworld structure into latent space.** The embeddings of relevant variables are compressed into interactions which are then expanded to a discrete probability distribution over possible realizations of the world. The full process consists of first *(i) compressing* embedding vectors to their scalar interactions. Then *(ii) expanding* pairwise interactions to pairs of vectors, *expanding* the vector pairs to matrices through an outer product, and (for $C \geq 3$) *expanding* the matrices into a tensor by broadcasting each matrix over all other dimensions. Finally, a valid *(iii) likelihood* is ensured by a sigmoid transformation for normalization.

A.3 Experiment 1:

Network goal belief. The goal belief B^{Net}_{tgr} is extracted from the accumulated output $\mathcal{M}_{tcr} = \sum_{t'=1}^{t} M_{t'cr}$ (where $M_{t'cr}$ are the single-timestep network pre-activations) via a softmax over possible realizations $B^{\text{Net}}_{tgr} = \frac{e^{\mathcal{M}_{tgr}}}{\sum_{r'} e^{\mathcal{M}_{tgr'}}}$. Although observations are conditionally independent given the realizations $\mathbf{r}$, the optimal estimate of r_g requires marginalizing the joint likelihood over the non-goal realizations $\mathbf{r}_{\setminus g}$ before maximizing:

$$\arg\max_{r_g} P_Z(r_g \mid \mathbf{o}_{1:T}) = \arg\max_{r_g} \left(\sum_{\mathbf{r}_{\setminus g}} \exp\left(\sum_{t=1}^{T} \ln P_Z(\mathbf{o}_t \mid \mathbf{r}) \right) \right).$$

Crucially, because the summation over $\mathbf{r}_{\setminus g}$ envelops the temporal accumulation, the log-likelihood cannot be decomposed into a simple iterative summation over time for the marginals. Therefore, sequential Bayesian updating is only Markovian with respect to an intractable *joint* posterior $P_Z(\mathbf{r} \mid \mathbf{o}_{1:t})$. Since ΔB_{tc} is restricted to marginal beliefs to remain tractable, re-evaluating the

marginal B_{tc} in light of a new observation $\mathbf{o}_{t+1}$ will not be Markovian. Therefore, performing this update accurately requires the network's memory state to encode the missing joint dependencies—specifically, the interaction information—that is lost when only tracking marginals.

Classifier loss. To provide a rich teaching signal, we minimize a symmetrized KL divergence between the network and optimal goal belief-state (where $\alpha = 0.5$ gave the best results):

$$\mathcal{L}_{\text{FR}} = \left(\frac{1}{2} \mathcal{D}_{\text{KL}}(B_{tg}^{\text{Joint}} \,\|\, B_{tg}^{\text{Net}}) + \frac{1}{2} \mathcal{D}_{\text{KL}}(B_{tg}^{\text{Net}} \,\|\, B_{tg}^{\text{Joint}}) \right)^{\alpha}.$$

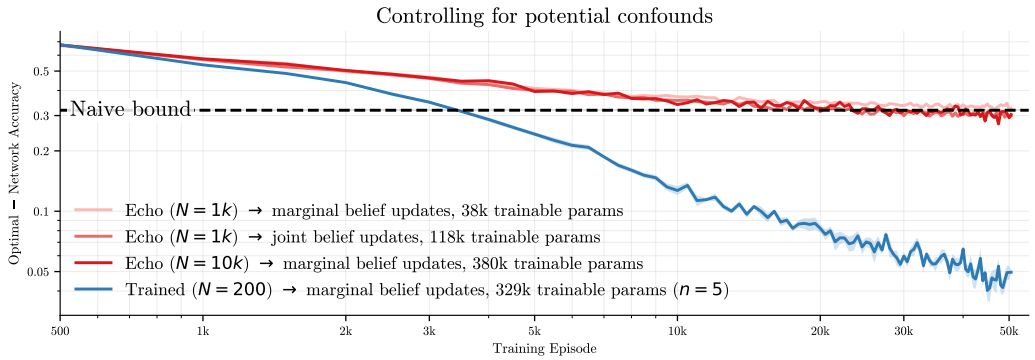

Figure 9: **Results survive additional controls** ($C = 2$)**.** Excess error above optimality at the final step of inference, plotted against training episode on log–log axes. The mean ($\pm$ SEM) training curve of 5 small Fully Trained RNNs is shown in blue. ESNs are colored light to solid red by their number of trainable parameters, where $N = 10{,}000$ matches the Fully Trained RNN with $N = 200$ and another ESN ($N = 1{,}000$) outputs updates to the joint belief-state.

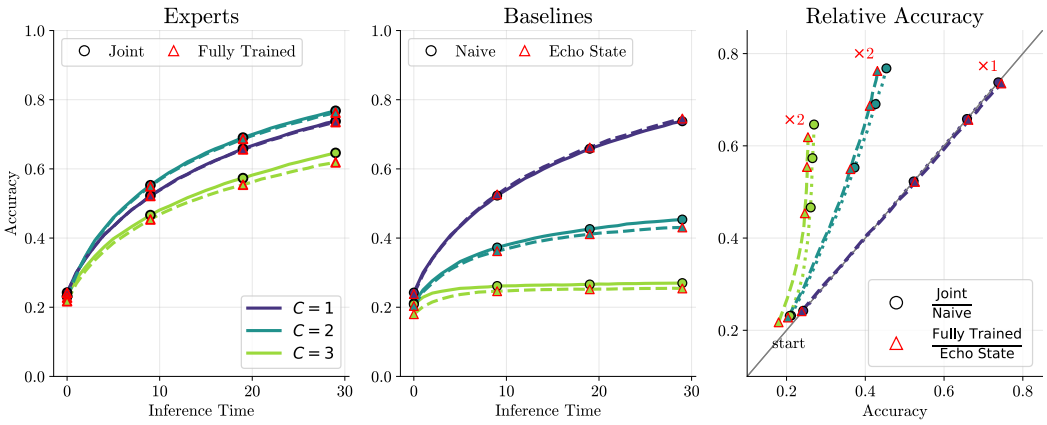

Figure 10: **Results extend to 3 interacting variables.** Same as Figure 3 with $C = 3$ ($n = 1$).

Confidence–accuracy calibration. We define an agent's confidence as the belief it assigns to its selected action, B_{Tga} with $a = \arg\max_r B_{Tgr}$, and its accuracy as the fraction of episodes in which that action was correct. Figure 11a fits accuracy linearly against confidence for $C = 1, 2, 3$: Joint Bayes and the Fully Trained RNN lie on the identity line, whereas Naive Bayes and the ESN are increasingly over-confident as C grows. Figure 11b splits episodes by Factorization Regret and plots the ratio of accuracy to confidence. The degree of misalignment grows with the number of interacting variables in agents that neglect Factorization Regret.

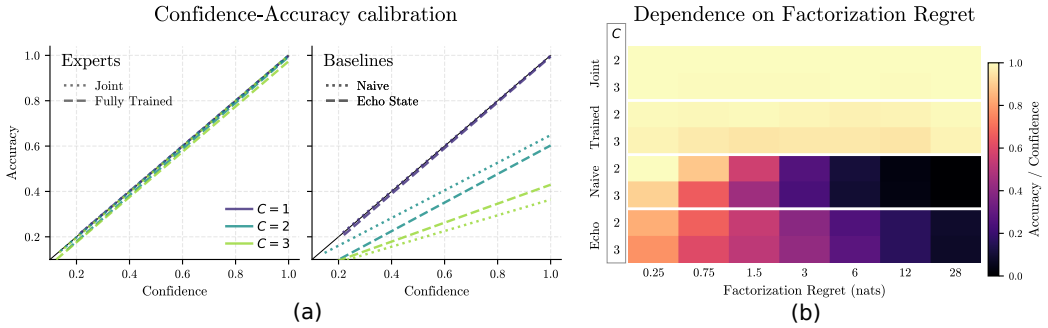

Figure 11: **Confidence–accuracy calibration for** $C = 1, 2, 3$**. (a)** Linear fits between confidence in the selected action and accuracy. **(b)** Accuracy over confidence as a function of Factorization Regret.

A.4 Experiment 2:

Classifier loss. At the end of each trajectory we sample an action $a \sim B_{Tgr}$ and receive outcome $\delta(a - r_g) \in \{0, 1\}$. We train using a regularized cross-entropy-style objective:

$$\mathcal{L}_{\text{Classifier}} = \frac{-1}{T} \sum_t \Big(\underbrace{\delta(a - r_g)}_{\text{rewarded}} \ln B_{tga} + \underbrace{\big(1 - \delta(a - r_g)\big)}_{\text{unrewarded}} \ln(1 - B_{tga}) - \beta_1 \underbrace{B_{tga} \ln B_{tga}}_{\text{regularization}} \Big),$$

Gradients from $\mathcal{L}_{\text{Classifier}}$ update only the Classifier parameters. Actions are sampled ($a \sim B_{Tgr}$) during training; at evaluation, the action is taken deterministically as $a = \arg\max_r B_{Tgr}$.

To provide the Generator estimates of the relevant variable realizations we sample $\hat{\mathbf{r}}$ from the Classifier's final belief-state B_T. We assume that the Classifier has already acted and insert the taken action a into the g'th index of $\hat{\mathbf{r}}$. We also provide the Classifier's confidence in the sampled realizations $\hat{\mathbf{r}}_{conf} = B_{T\hat{\mathbf{r}}}$. The g'th index of $\hat{\mathbf{r}}_{conf}$ is set to 1 if a was rewarded, 0 otherwise.

Generator loss. We sample a latent configuration $\hat{\mathbf{r}}$ from the Classifier belief at $t = T$ (*without* a gradient), and train the Generator to maximize the likelihood of observations. Since the true generative distribution is unknown, we construct an empirical target distribution $\tilde{P}(\mathbf{o})$, where the likelihood of each independent observable is given by its trajectory average $\tilde{P}_Z(o^i = 1) = \langle o^i \rangle_t$. We use this to compute a self-supervised (since the target is constructed from observations) Jeffreys Divergence loss:

$$\mathcal{L}_{\text{Generator}} = J\Big(\tilde{P}(\mathbf{o}) \,\|\, \hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}}) \Big).$$

The gradient from $\mathcal{L}_{\text{Generator}}$ updates the agent's Compositional Embeddings $\{\hat{\mathbf{k}}_c^i, \hat{\mathbf{q}}_c^i\}_{i=1}^{d_o} \in \hat{\mathcal{E}}, \quad \forall c \in \mathcal{C}$ via gradient flow through $\hat{Z}$.

Training reliability was improved by augmenting the Generator's loss function to give greater weight to elements of a batch where the Classifier's action was rewarded. The augmented loss was:

$$\mathcal{L}^{\star}_{\text{Generator}} = \frac{\text{chance}}{d_o} \times \mathcal{L}_{\text{Generator}} + \frac{1 - \text{chance}}{\langle \delta(a - r_g) \rangle} \times \mathcal{L}^{\delta}_{\text{Generator}} + \mathcal{L}_{\text{reg}},$$

where chance $= \frac{1}{R}$, $\mathcal{L}_{\text{Generator}}$ is averaged over all elements of the batch, $\mathcal{L}^{\delta}_{\text{Generator}}$ is the same average with unrewarded elements set to zero (so that the second term averages over rewarded elements only), and $\langle \delta(a - r_g) \rangle$ is the average accuracy in the batch. The factor $1/d_o$ reflects that $\mathcal{L}_{\text{Generator}}$ sums the divergence over observables, so the first term is an average over observables. All Experiment 2 results use $\mathcal{L}^{\star}_{\text{Generator}}$.

Gradient stabilization (soft clip). To avoid vanishing gradients when clipping probabilities, we use a differentiable soft clip:

$$\text{soft clip}(x, \epsilon_{clip}) = \epsilon_{clip} + \sigma\left(\frac{\hat{x} - \epsilon_{clip}}{\epsilon_{clip}} \right) (\hat{x} - \epsilon_{clip}),$$

$$\hat{x} = (1 - \epsilon_{clip}) - \sigma\left(\frac{1 - \epsilon_{clip} - x}{\epsilon_{clip}}\right)(1 - \epsilon_{clip} - x).$$

We found training reliability to improve dramatically when $\epsilon_{clip} = 10^{-3}$. The soft clip is applied during training to the Classifier's marginal beliefs and to the probabilities entering each divergence.

Learnable Compositional Embeddings. Training reliability was improved when the learnable embedding estimates were over-parameterized. The relevant variable embedding estimates were passed through learnable linear layers $f_K, f_Q : \mathbb{R}^{d_{\hat{\mathcal{E}}}} \to \mathbb{R}^{d_{\mathcal{E}}}$, to project them down from $d_{\hat{\mathcal{E}}} = N$ to dimensionality $d_{\mathcal{E}}$. Estimated interactions were computed from the projections:

$$\hat{\mathbf{z}}^i = \langle f_K(\hat{K}^i), f_Q(\hat{Q}^i) \rangle$$

Embedding regularization was applied to the projections:

$$\mathcal{L}_{\text{reg}} = \frac{1}{C \times d_o} \sum_{c,i} \left[(\|f_K(\hat{\mathbf{k}}_c^i)\|_2 - 1)^2 + (\|f_Q(\hat{\mathbf{q}}_c^i)\|_2 - 1)^2 \right]$$

Generalization emerges at a critical number of latent variables in set $\mathcal{S}$. When $\hat{Z}$ was learned, we found that varying the total number of latent variables in the environment had a significant impact on generalization (Figure 12). When the environment contained too few latent variables ($|\mathcal{S}| < 500$), only training variables are learned. As the number of latent variables in $\mathcal{S}$ increased, learning on the training set slowed, until spontaneously recovering and generalizing to testing variables.

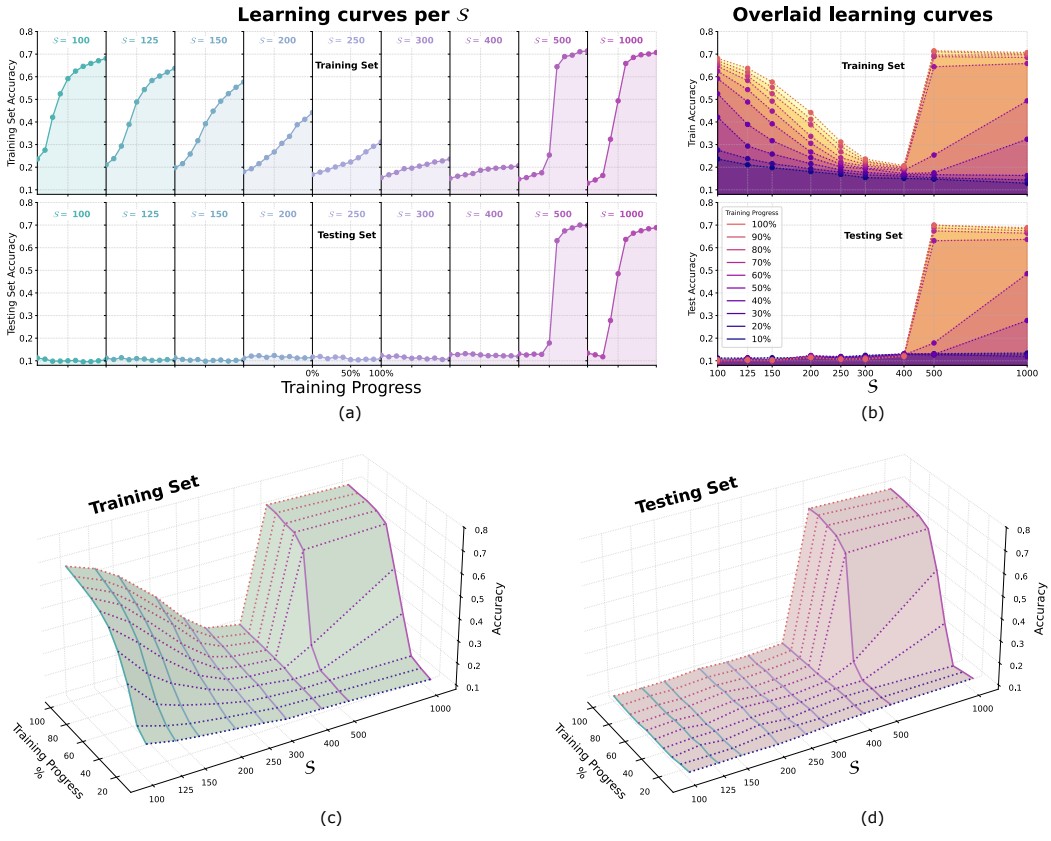

Figure 12: **Learning and generalization with Compositional Embeddings depend on $|\mathcal{S}|$ ($C = 2$). (a)** Learning dynamics for training (top) and testing (bottom) variables, split by the number of latent variables $|\mathcal{S}|$. **(b)** All sizes of $\mathcal{S}$ overlaid and split by training progress. **(c-d)** 3-dimensional visualization of training and testing set accuracy as a function of $|\mathcal{S}|$ and training progress.

Generator ablation impairs performance. To isolate the contribution of the gradient from the Generator we trained RCCs in a simplified setting that removed all other challenges for learning. In both forward and backward passes, the Classifier's sampled actions were replaced by the ground truth realizations $\mathbf{r}$, providing an exact teaching signal to the Classifier and a reliable input to the Generator. Additionally, we used an orthonormal initialization of f_K, f_Q with gain $\sqrt{N/d_{\mathcal{E}}}$ and zero bias to capture the statistics of $\mathcal{E}$. To further facilitate learning, we implemented layer normalization, skip connections and initialized the input layers of $\hat{Z}$ to be aligned with the readouts. Figure 13 demonstrates that RCC training is far more reliable when the gradient to $\hat{\mathcal{E}}$ flows from the Generator rather than from the Classifier.

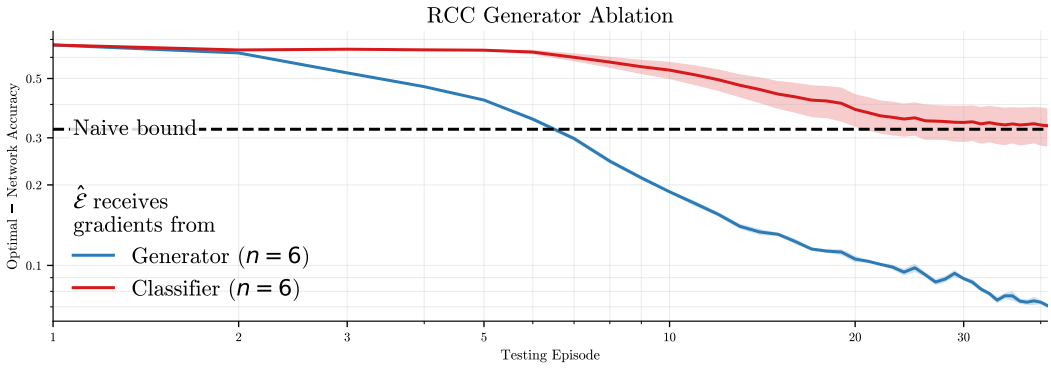

Figure 13: **Learning Compositional Embeddings through the classifier is unreliable** ($C = 2$). Excess error above optimality at the final step of inference (mean ± SEM with 6 seeds), plotted against testing episode (1 test per 500 training episodes) on log–log axes. The embeddings receive gradients from either the Generator (blue) or Classifier (red).

A.5 Experiment 3:

Controller loss. The Controller is trained using an Advantage Actor-Critic (A2C) algorithm to maximize the intrinsic reward:

$$\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell) = \exp\left(\frac{1}{d_o}\sum_i \left(\ln(\ell_{\mathbf{z}^i}(\mathbf{r}))\Omega^i + \ln(1-\ell_{\mathbf{z}^i}(\mathbf{r}))(1-\Omega^i)\right)\right), \tag{1}$$

The Actor and Critic both receive $\hat{Z}$, and output a policy over $\mathbf{r}$ and the expected $\mathcal{R}$, respectively. A single preferred observation $\mathbf{\Omega}$ is fixed for each Controller run, whereas the context, and hence $\hat{Z}$, varies across the elements of a batch; in the Offline conditions the same set of contexts is used in every episode. At each episode, $\mathbf{r}$ is sampled stochastically from the Actor's policy $\pi(\mathbf{r} \mid \hat{Z})$. In contrast to the Classifier, whose outputs are marginal beliefs, the Actor's policy is a probability distribution over the full action space so $\mathbf{r}$ is sampled from all dimensions jointly. The Critic predicts the intrinsic reward $\mathcal{R}$. The advantage A is calculated as the difference between the true intrinsic reward and the Critic's prediction:

$$A = \mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell) - \nu(\hat{Z})$$

The Controller loss $\mathcal{L}_{A2C}$ integrates the Critic's prediction error, the Actor's policy gradient and a policy entropy bonus weighted by β_{cntrl} to encourage exploration:

$$\mathcal{L}_{A2C} = A^2 - A_{\text{Detached}} \ln(\pi(\mathbf{r} \mid \hat{Z})) + \beta_{cntrl} \sum_{\mathbf{r}'} \pi(\mathbf{r}' \mid \hat{Z}) \ln \pi(\mathbf{r}' \mid \hat{Z}).$$

Controller training comparisons. The schematic in Figure 14a demonstrates the mapping from preferences to specific likelihoods (and their cumulative product) into a *preference landscape* of $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$. Figure 14b shows mental navigation of the resultant preference landscape. We compare Controller training across three conditions that vary in how they estimate $\hat{Z}$ and ℓ:

In the *Online condition* **(i)**, the Controller directly interacts with the environment, setting the state of the world to a joint realization sampled from the Actor's policy. $\hat{Z}$ is learned from scratch by training a Generator (like in subsection A.4, but since this agent directly controls the world there is no uncertainty around $\mathbf{r}$ and therefore $\hat{\mathcal{E}}$ is learned much faster). Because every sampled $\mathbf{r}$ is correct, this Generator is trained without the reward weighting of $\mathcal{L}^{\star}_{\text{Generator}}$ (its divergence is averaged uniformly over the batch and observables), and its confidence input is fixed at 1. The likelihoods ℓ in $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$ are approximated using the empirical average observation $\langle \mathbf{o} \rangle_t$ of the episode.

The other two agents learn *Offline*, sampling a joint realization $\mathbf{r}$ and imagining the resulting observations, without any manipulation of the environment. These agents receive an estimate of $\hat{Z}$ coming from embedding space $\hat{\mathcal{E}}$ which was passively learned through Representation Classification Chains (an RCC with the architecture and training procedure of subsection A.4, trained separately for this experiment with the hyperparameters listed in subsection B.5). In order to approximate ℓ in $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$, the *Offline agent w/ Generator* **(ii)** uses the Generator that was pre-trained by the RCC to approximate $\ell_{\hat{\mathbf{z}}^i}(\mathbf{r}) \leftarrow \hat{P}_{\hat{\mathbf{z}}^i}(o^i = 1 \mid \mathbf{r})$. Serving as an oracle for comparison, the *Offline agent w/ Joint Bayes* **(iii)** uses the ground-truth $\ell_{\mathbf{z}^i}(\mathbf{r}) \leftarrow P_{\mathbf{z}^i}(o^i = 1 \mid \mathbf{r})$.

Training fully offline with the generative model achieves similar performance and efficiency to training offline with direct access to the joint likelihood (Figure 14c). Training online with direct interactions with the environment is no better than training offline. Training was performed with a batch of 20,000 trajectories per episode. In the Online condition, all 20,000 trajectories were followed by causal manipulations of the environment prior to the onset of the next episode.

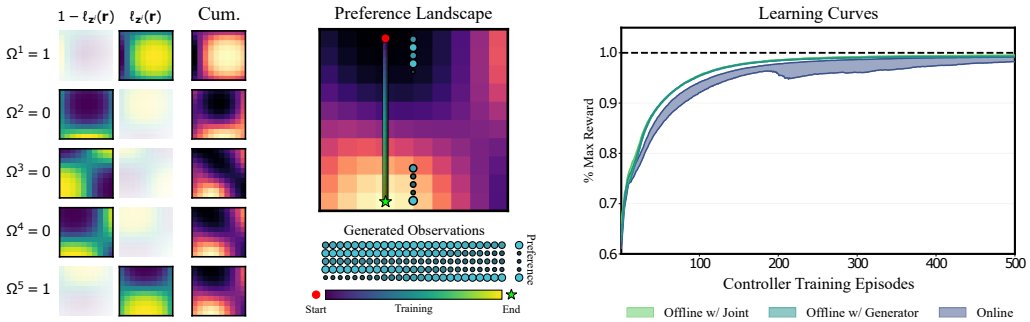

Figure 14: **Conditional generative modeling enables optimization in compositional spaces. (a)** Schematic illustrating the mapping of preferred observations ($\mathbf{\Omega}$) to their respective likelihoods and the cumulative landscape (accumulated over i in Equation 1). **(b)** An example traversal, from the lowest to the highest point on the landscape, changes observations to best match the agent's preference. **(c)** Controller learning curves (mean $\pm$ SEM, $n = 20$ initializations of $\mathbf{\Omega}$). % max reward is calculated as $\mathcal{R}(\arg\max_{\mathbf{r}} \pi(\mathbf{r} \mid \hat{Z}), \mathbf{\Omega}, \ell) \quad / \quad \max_{\mathbf{r}} \mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$, using the ground truth ℓ.

Relation to offline model-based optimization. Offline model-based optimization seeks to identify the input that maximizes an unknown objective given only a static dataset of inputs and their scores; a model of the objective is fit to these data and then optimized, which tends to produce inputs on which the model erroneously overestimates the objective (Trabucco et al., 2021). Experiment 3 shares this structure, but the preference $\mathbf{\Omega}$ is internal and known, whereas the generative process ℓ had to be estimated and no objectively scored inputs exist. Optimizing against $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \mathbf{r})$ exposes the Controller to the same failure mode as offline model-based optimization and policies trained inside learned world models (Ha & Schmidhuber, 2018). Because all conditions are evaluated under the ground-truth ℓ (Figure 14c), such a failure would appear as a gap between the Offline agent w/ Generator and the oracle.

Choice of learner. While Reinforcement Learning is not strictly necessary for solving this control problem (the joint action space $R^C = 100$ was small enough to enumerate), our goal was to demonstrate that simple memory-less RL agents *can* solve this problem. Modern memory-less RL

algorithms could therefore be employed in more challenging settings where enumeration becomes impractical (as C grows).

B SUPPLEMENTARY

Important properties of likelihood constructions. We chose our manner of expansion to ensure several properties:

- **Low rank structure:** Broadcasting $\mathbf{v}^i_{cc'}(\mathbf{v}^i_{c'c})^\top \in \mathbb{R}^{R\times R}$ to $\mathbb{R}^{R^C}$ constructs the likelihood entirely from pairwise interactions, embedding low-rank structure into the tensor.
- **Constrained scale:** Using a sine function for $\mathbf{v}^i_{cc'}$ creates a linear (as a function of C) bound on the standard deviation of (pre-activation) elements of ℓ.
- **Visual Interpretability:** Constructing $\mathbf{v}^i_{cc'}$ with a smooth function ensures local similarity in $P_Z(\mathbf{o} \mid \mathbf{r})$ such that similar vectors $\mathbf{r}$ generate similar observations.
- **Consistency:** Using $z^i_{cc'}$ to modulate phase, rather than scale, ensures different contexts and dimensions of $\mathcal{O}$ generate similar amounts of information.
- **Useful Baseline:** Ensuring the potential for asymmetry in $\mathbf{v}^i_{cc'}$ ensures that naive inference can perform above chance.

Future Directions of the Generative Process. A natural extrapolation would be to construct $\mathbf{v}^i_{cc'}$ using Fourier series, perhaps adding another dimension to K and Q for frequency specific interaction. Such a design would embed structure at multiple spatial frequencies while staying interpretable, similar to modules of grid cells in the medial entorhinal cortex (Burak, 2014). Another natural extension would be to introduce sparsity to K and Q, to move away from the current regime where all dimensions of $\mathcal{E}$ have *some degree* of interaction. In summary, two promising future directions:

- Construct ℓ through inverse discrete Fourier transform (rather than a single frequency) for greater expressivity.
- Sparsify embeddings by masking elements in $\mathcal{E}$ for greater interaction specificity.

B.1 Factorization Regret Relation to Interaction Information

For clarity we first focus on the Interaction Information between a single observable o^i and two random variables, $I(r_g; r_c; o^i)$; the results extend to the full observation $\mathbf{o}$, whose channels are independent given $\mathbf{r}$. Since r_g and r_c are a priori independent in our setting, we can write:

$$I(r_g; r_c; o^i) = -I(r_g; r_c \mid o^i) = -\mathbb{E}[\mathcal{D}_{\mathrm{KL}}(P_{\mathbf{z}^i}(r_g, r_c \mid o^i) \,||\, P_{\mathbf{z}^i}(r_g \mid o^i) \otimes P_{\mathbf{z}^i}(r_c \mid o^i))]$$

Now consider two Bayesian observers processing a sequence of observations $o^i_{1:T}$:

1. **Optimal Observer**: Updates its joint belief $B_t(r_g, r_c)$ correctly using $P_{\mathbf{z}^i}(o^i | r_g, r_c)$.
2. **Naive Observer**: Updates its joint belief $\tilde{B}_t(r_g, r_c)$ using factorized likelihoods

$$\tilde{P}_{\mathbf{z}^i}(o^i | r_g, r_c) \propto P_{\mathbf{z}^i}(o^i \mid r_g) \otimes P_{\mathbf{z}^i}(o^i \mid r_c).$$

Since the Naive observer has separate likelihoods for r_g and r_c, its marginal beliefs are independent:

$$\tilde{B}_t(r_g \mid o^i_{1:t}) \;\propto\; p(r_g) \prod_{t' \le t} P_{\mathbf{z}^i}(o^i_{t'} \mid r_g) \underbrace{\sum_{r_c} p(r_c) \prod_{t' \le t} P_{\mathbf{z}^i}(o^i_{t'} \mid r_c)}_{\text{independent of } r_g},$$

and it treats observations as i.i.d. given a single latent variable.

Interaction Information emerges in the joint posterior before impacting marginal beliefs. For a *single observation (o_1)*, the Optimal observer's marginal belief:

$$B_1(r_g \mid o^i_1) \propto p(r_g) \sum_{r_c} p(r_c) P_{\mathbf{z}^i}(o^i_1 \mid r_g, r_c)$$

coincides exactly with the Naive observer's:

$$\tilde{B}_1(r_g \mid o_1^i) \;\propto\; p(r_g) P_{\mathbf{z}^i}(o_1^i \mid r_g).$$

Yet a divergence can already exist between their joint posteriors, quantified by $I(r_g; r_c; o_1^i)$:

$$I(r_g; r_c; o_1^i) = -\mathbb{E}[\mathcal{D}_{\mathrm{KL}}(B_1(r_g, r_c|o_1^i) \,||\, \tilde{B}_1(r_g|o_1^i) \otimes \tilde{B}_1(r_c|o_1^i))]$$

Thus, for the first observation, the divergence between Optimal and Naive posteriors can contain Interaction Information, despite full agreement between their marginal beliefs. Critically, this Interaction Information plants the seed for the marginal beliefs to diverge on the next step of inference.

Factorization Regret measures the accumulating impact on marginal beliefs. The joint posteriors of the first step become the joint priors for the second. Any divergence between these joint priors allows *subsequent observations* to create divergence in marginal beliefs. *Factorization Regret (FR)* quantifies this accumulating divergence:

$$\mathbb{E}[\text{Factorization Regret}_T] = \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(B_T(r_g|\mathbf{o}_{1:T}) \,||\, \tilde{B}_T(r_g|\mathbf{o}_{1:T}))].$$

This consists of the conditional Total Correlation (TC) minus an offset. Applying the chain rule for $\mathcal{D}_{\mathrm{KL}}$ to the joint $P_Z(r_g, \mathbf{o}_{1:T})$ against its factorized counterpart $p(r_g)\prod_t P_Z(\mathbf{o}_t \mid r_g)$:

$$\mathbb{E}[\text{Factorization Regret}_T] = \underbrace{\mathbb{E}\Big[\mathcal{D}_{\mathrm{KL}}\big(P_Z(\mathbf{o}_{1:T}|r_g) \parallel \textstyle\prod_t P_Z(\mathbf{o}_t|r_g)\big)\Big]}_{:= TC(\mathbf{o}_{1:T}|r_g)} - \mathcal{D}_{\mathrm{KL}}\big(P_Z(\mathbf{o}_{1:T}) \parallel \tilde{P}_Z(\mathbf{o}_{1:T})\big),$$

where $\tilde{P}_Z(\mathbf{o}_{1:T}) = \sum_{r_g} p(r_g) \prod_t P_Z(\mathbf{o}_t \mid r_g)$. The term $TC(\mathbf{o}_{1:T} \mid r_g)$ expresses the reward-relevant information contained in the trajectory that is lost by a myopic focus on a goal.

Interactions in latent space generate temporal dependence between observations. The conditional total correlation $TC(\mathbf{o}_{1:T} \mid r_g)$ is a statement about *time*, while $I(r_g; r_c; \mathbf{o})$ is a statement about *latent space*. Because the realizations $\mathbf{r}$ are drawn once per episode and observations are i.i.d. given $\mathbf{r}$, the full set of latent variables explains all temporal dependencies:

$$TC(\mathbf{o}_{1:T} \mid r_g) \;=\; \underbrace{TC(\mathbf{o}_{1:T} \mid r_g, r_c)}_{=0} \;+\; \sum_t I(\mathbf{o}_t; r_c \mid r_g) \;-\; I(\mathbf{o}_{1:T}; r_c \mid r_g)$$

Because sequential inference is permutation invariant we can consider a generic observation $\mathbf{o}$:

$$TC(\mathbf{o}_{1:T} \mid r_g) \;=\; T \times (I(\mathbf{o}; r_c) - I(\mathbf{o}; r_c; r_g)) \;-\; I(\mathbf{o}_{1:T}; r_c \mid r_g).$$

The first term is positive because $I(\mathbf{o}; r_c; r_g) \leq 0$ and the second term is bounded by the entropy of r_c. Thus, the total correlation may grow linearly with time, with Interaction Information $I(\mathbf{o}; r_c; r_g) \leq 0$ determining the average rate of growth. This reveals a direct mapping from the geometry in latent space to the temporal dependencies in a trajectory.

Relation to Semantic Information. Kolchinsky & Wolpert (2018) define semantic information as the information a system has about its environment that is causally necessary for maintaining its *viability* (i.e. existence), and identify it through interventions that scramble system–environment correlations. The Naive observer applies such an intervention to Bayesian inference: it replaces the joint likelihood with the product of its marginals, scrambling the Interaction Information while preserving each variable's marginal likelihood. Factorization Regret measures the consequence of this intervention for the goal, with the belief over the optimal action in the role of viability. It is therefore the analogue of their value of information under full scrambling. The single-observation case above makes the distinction concrete: Interaction Information is already present in the joint posterior, yet Factorization Regret is zero because the belief over the goal is unaffected. Interaction Information thus becomes semantic only once it bears on the goal.

B.2 Learning Dynamics

Average performance improves throughout training for both Fully Trained RNNs and ESNs (Figure 15a-b). However, we hypothesize that the ESN is immediately on a trajectory towards a suboptimal *stimulus-bound* strategy (Vertechi et al., 2020). To test this hypothesis, we investigate the

Participation Ratio (i.e. effective dimensionality) of the Read-in and recurrent activity (Figure 15c). While a Fully Trained network expands the dimensionality of its recurrent dynamics relative to the Read-in, the ESN shows the opposite effect. Although the ESN accuracy improves, a large input gain, relative to the recurrent dynamics, could make it difficult to learn when subtle changes in a trajectory correspond to large jumps in latent space.

We next investigated whether Factorization Regret could reveal the strategies being learned. Early into training a negative correlation emerges between accuracy and Factorization Regret across the batch of trajectories (Figure 15d). Since the networks initially know nothing, the correlation between FR and accuracy begins at zero. Failure rapidly becomes predictable from FR; however, this correlation dissolves for the Fully Trained RNN and persists in the ESN. To demonstrate that this is not a trivial outcome of performance differences, we plot this correlation against the average accuracy in the episode (Figure 15e), demonstrating that even when networks have the same accuracy on average, they are developing different strategies with respect to FR. This capacity to mitigate the correlation between FR and accuracy appears to coincide with an increase in recurrent dimensionality on average (Figure 15f).

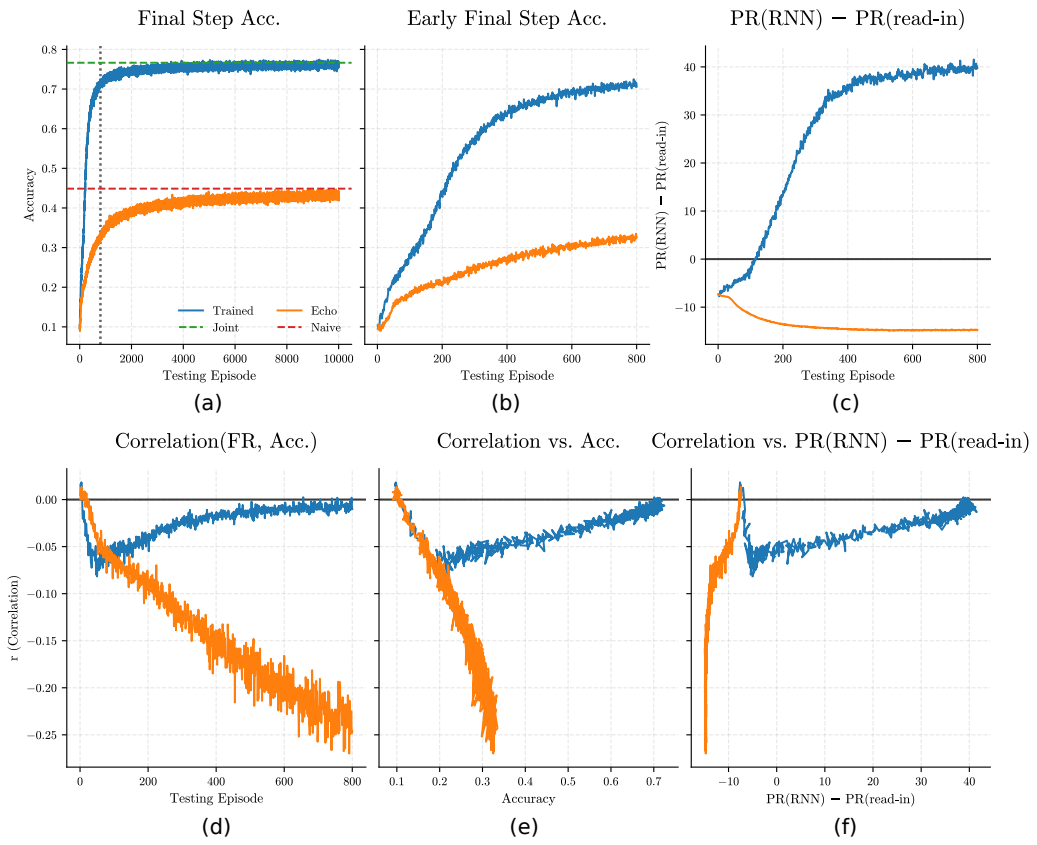

Figure 15: **Managing Factorization Regret involves high-dimensional RNN dynamics** ($C = 2$)**.** **(a)** Throughout learning, testing accuracy at the final step of inference saturates to the performance of either Joint or Naive Bayes. The vertical dotted line marks the range shown in (b–f). **(b)** Final step testing accuracy during early training. **(c)** Difference in Participation Ratio (PR) between the recurrent and read-in activity. **(d)** Correlation between Factorization Regret and accuracy at the final step of inference. A negative correlation emerges in the beginning of training and persists for the ESN. **(e)** Correlation versus accuracy. **(f)** Correlation versus PR difference.

B.3 Entanglement

Misalignment between the ESN's internal space and the true latent space could be explained mechanistically as entanglement. Entanglement in an RNN's representation may emerge from the mixing

of information by temporal correlations between neurons (Miller et al., 2023). Therefore, we hypothesized that training only a read-in and read-out on an evidence accumulation task, which demands temporal correlations, would not only fail to overcome this entanglement, but would exacerbate it. Indeed, Principal Component Analysis (PCA) on the marginal beliefs for network and Bayesian agents was consistent with this prediction (Figure 16). When performing PCA on marginal beliefs after a single observation, all representations appeared factorized. Yet PCA on the marginal beliefs at the end of inference appeared entangled for the ESN and Naive Bayes which lose the ability to differentiate r_c from $r_{c'}$. Additionally, the Joint and Fully Trained representations allocate low entropy beliefs (i.e. high confidence) to all $(r_c, r_{c'})$ pairs. In contrast, the ESN and Naive Bayes represent confidence orthogonally to r_c and $r_{c'}$.

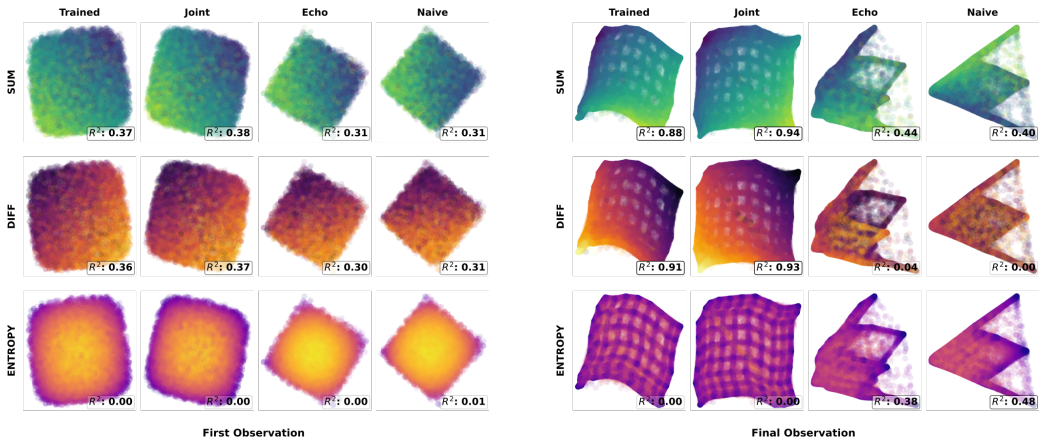

Figure 16: **Factorization Regret leads to entanglement over time.** Top 2 Principal Components of the marginal beliefs after only a single observation (left) and $T = 30$ observations (right). Here c and c' denote the two variables in context ($C = 2$), one of which is the goal. Beliefs are colored by the realization sum ($r_c + r_{c'}$), difference ($r_c - r_{c'}$), and belief entropy ($-\sum_c \sum_r B_{tcr} \ln B_{tcr}$). R^2 indicates the variance explained of the respective variables by the top 2 components. After only a single observation, representations are still perfectly factorized. By the final observation, ESN and Naive representations become deeply entangled, while the Fully Trained representation remains factorized alongside Optimal Joint Bayes.

For a mechanistic account of this entanglement, let $\mathbf{e}_r \in \{0,1\}^R$ represent a one-hot indicator vector which inquires about the probability of realization r. In $P_{z^i}(o^i \mid r)$, the mapping from scalar z^i to a vector of R updates is mediated by the pre-sigmoid expansion $f(z^i)$ (the vector $\mathbf{v}^i$ of subsection A.2). For $C = 1$, inquiring about a specific realization $r = r'$ projects the indicator vector onto the likelihood via:

$$P_{z^i}(o^i = 1 \mid r = r') = \sigma\big(\mathbf{e}_{r'}^T f(z^i)\big).$$

Crucially, because **different realizations of the same variable are mutually exclusive**, the likelihood of a specific $r = r'$ is perfectly isolated from all other potential realizations and $P_Z(r = r' \mid \mathbf{o})$ may be updated independently. Next, consider a context of two latent variables ($C = 2$). Here the interactions $Z^i = (z^i_{gc}, z^i_{cg})$ map to a joint space via the bilinear tensor product $f(z^i_{gc}) \otimes f(z^i_{cg}) \in \mathbb{R}^{R \times R}$, forming an $R \times R$ matrix of belief updates. The bilinear construction generates synergy. In other words, the combination of variables loses predictive power if the variables are considered in isolation, i.e. the whole is greater than the sum of its parts (Proca et al., 2024; Barak et al., 2013). We can see this because evaluating the likelihood of $o^i = 1$ given a realization $r_g = r'$ now involves projecting indicator $\mathbf{e}_{r_c}$ onto a basis shared with $\mathbf{e}_{r_g}$:

$$P_{Z^i}(o^i = 1 \mid r_g = r', r_c) = \sigma\Big(\mathbf{e}_{r'}^T \left(f(z^i_{gc}) f(z^i_{cg})^T\right) \mathbf{e}_{r_c}\Big).$$

In summary, the entanglement between latent variables emerges from projecting onto a shared basis.

Next we turn to investigating entanglement in real-time. Since FR is an accumulating measure (best used for comparing full trajectories) it is not well suited for high temporal fidelity analysis. Instead,

we aim to compare the Naive marginal likelihood $P_Z(\mathbf{o}_t \mid r_g)$ and the history-conditioned marginal likelihood $P_Z(\mathbf{o}_t \mid r_g, \mathbf{o}_{1:t-1})$. The history-conditioned marginal likelihood is proportional in r_g to the ratio of the Optimal Observer's successive marginalized beliefs, since the evidence term $P_Z(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})$ does not depend on r_g:

$$P_Z(\mathbf{o}_t \mid r_g, \mathbf{o}_{1:t-1}) = \sum_{r_c} P_Z(\mathbf{o}_t \mid r_g, r_c) p(r_c \mid r_g, \mathbf{o}_{1:t-1}) \propto \frac{B_t(r_g)}{B_{t-1}(r_g)}.$$

The history-dependence of the Optimal Bayesian update can be attributed to the prior's impact on the functional likelihood, $\sum_{r_c} P_Z(\mathbf{o}_t \mid r_g, r_c) p(r_c \mid r_g, \mathbf{o}_{1:t-1})$, manifesting as entanglement between variables r_g and r_c. To quantify this *"non-Markovian-ness"*, we measure the Jeffreys Divergence between (normalized) Naive marginal likelihoods and history-conditioned marginal likelihoods:

$$\text{Entanglement}_t = J\Big(\frac{P_Z(\mathbf{o}_t \mid r_g)}{\eta_a} \;||\; \frac{P_Z(\mathbf{o}_t \mid r_g, \mathbf{o}_{1:t-1})}{\eta_b}\Big) = J\Big(\frac{P_Z(\mathbf{o}_t \mid r_g)}{\eta_a} \;||\; \frac{B_t(r_g)}{B_{t-1}(r_g)\,\eta_b}\Big),$$

where η_a and η_b are normalization factors over r_g to ensure valid probabilities; normalization removes the r_g-independent evidence term, so the two expressions are equal.

To investigate the response of the network dynamics to entanglement, we performed a cross correlation between Optimal Bayesian belief Entanglement_t and various properties of the network output. Figure 17a demonstrates that entanglement is associated with a change of activity in population space, in both Fully Trained RNNs and ESNs. However, the response differs greatly between networks. In the Fully Trained RNN, Entanglement is associated with an expansion of the population vector, whereas the ESN activity appears to shrink towards the origin. The Fully Trained RNN's expansion is also distributed throughout the population, rather than along the top component directions (Figure 17b). Interestingly, the top components of the Fully Trained RNN's activity are suppressed (Figure 17c), and the expansion is unexpectedly spread across many low variance components (Figure 17d). These dynamics could account for the increased RNN dimensionality observed in Figure 15c, indicating that dynamic control of network dimensionality may be a mechanism for maintaining a disentangled representation.

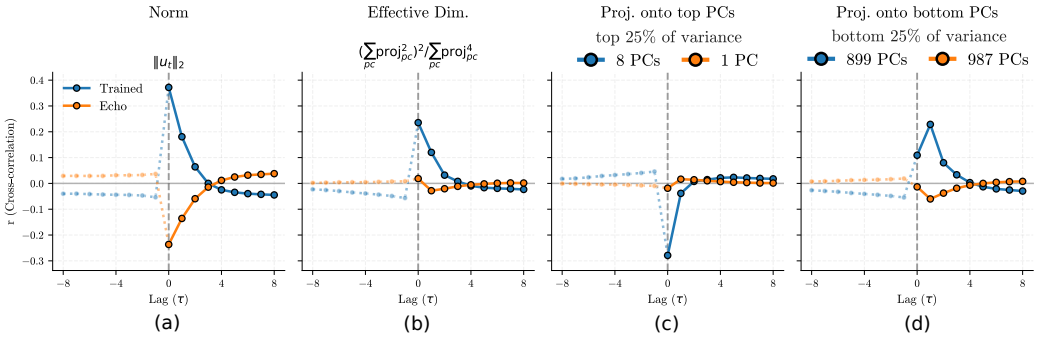

Figure 17: **Entanglement correlates with learned dynamics of dimensionality.** Cross-correlations between Bayesian entanglement and **(a)** the L2 norm of RNN output u_t, **(b)** the effective dimensionality of the projection onto the Principal Components, **(c)** the portion of the effective dimensionality explained by activity in the band of top Principal Components that explains 25% of neural variance and **(d)** same as (c) for the band of bottom Principal Components. Only data from $t > 20$ was analyzed in order to isolate the steady-state response profile. Legends give the number of Principal Components in each band.

B.4 Representation Classification Chains and Thousand Brains Theory

The Thousand Brains Theory of intelligence (Hawkins et al., 2018) postulates that individual cortical columns autonomously model through *compositions of compositions*. Tasked specifically with learning useful relationships, Representation Classification Chains can be viewed as the potential *microcircuit* of such a system. This architecture-agnostic circuit generates abstractions $\mathbf{r}$, to explain inputs $\mathbf{o}$, in a manner that enables optimal control of inputs given their internal value.

B.5 Additional Model Hyperparameters

All models are implemented in PyTorch.

General.

- Hidden dimensionality of all networks (N): 1000
- Parametric embedding dimensionality ($d_{\hat{\mathcal{E}}}$): 1000
- Optimizer: Adam

Experiment 1.

- Learning rate: 0.001
- Training episodes: 50000
- Training batch size: 8000
- Loss exponent (α): 0.5

Experiment 2.

- RCC Learning rate: 0.0005
- RCC Training episodes: 125000
- RCC Training batch size: 8000
- RCC Regularization weight (β_1): 0.1
- Testing: once per 100 training episodes

Experiment 3.

- RCC Learning rate: 0.0001
- RCC training episodes: 50000
- RCC Training batch size: 20000
- RCC Regularization weight (β_1): 0.01
- Controller Learning rate: 0.005
- Controller Training episodes: 500
- Controller Batch size: 20000
- Controller Entropy bonus (β_{cntrl}): 0.05
- Online Generator Learning Rate: 0.001

Classifier Architecture:

1. **Input:** At each time step t, the input x_t^{cls} is the concatenation of the current observation vector o_t and the interactions Z (or their estimate $\hat{Z}$ in Experiments 2–3): $x_t^{cls} = [o_t, Z]$.
2. **Read-in:** $h_t^{in} = \text{ReLU}(W_{in} x_t^{cls} + b_{in})$ maps the input to the hidden dimension N.
3. **LSTM:** $h_t^{stm}, h_t^{ltm} = \text{LSTM}(h_t^{in}, (h_{t-1}^{stm}, h_{t-1}^{ltm}))$ processes the $T \times N$ sequence where initial hidden states (h_0^{stm}, h_0^{ltm}) are learnable parameters.
4. **Read-out:** $l_t = W_{out} h_t^{stm} + b_{out}$ maps the LSTM output to C independent sets of R dimensional logits.
5. **Output:** As described in the main text, a cumulative sum is applied along time and then a softmax is applied along the realizations to produce the Classifier's belief-state.

The ESN uses the same architecture with its recurrent weights frozen at PyTorch's default initialization, with the read-in and read-out trained. Its reservoir therefore has $N = 1000$ units, with no spectral-radius rescaling or sparsification.

Generator Architecture:

Input: The Generator integrates three inputs: $x^{gen} = h_r + h_c + h_z, \quad$ where $h_r, h_c, h_z \in \mathbb{R}^N$.

- $h_r = W_s \,\mathrm{Emb}(\hat{\mathbf{r}}) + b_s$ is a linear projection of learnable PyTorch embeddings $\mathrm{Emb}(\hat{\mathbf{r}})$, concatenated across the C variables in context.
- $h_c = W_c \hat{\mathbf{r}}_{conf} + b_c$ is a linear projection of the confidence in the sampled realizations.
- $h_z = W_z \hat{\mathbf{z}}^i + b_z$ is a linear projection of estimated interactions $\hat{\mathbf{z}}^i$. Since dimensions of the observation space are independent we process them in parallel.

Output: $\hat{P}_{\hat{Z}}(\mathbf{o}|\hat{\mathbf{r}}) = \sigma\big(W_2(\mathrm{ReLU}(W_1(\mathrm{ReLU}(x^{gen})) + b_1)) + b_2\big)$.

Controller Architecture:

The Actor and Critic networks receive only $\hat{Z}$, and pass it through separate two-layer MLPs with ReLU activations. For the Actor, the final output is passed through a Softmax to produce a distribution over actions. For the Critic, the output is wrapped in a Sigmoid to predict the reward.

B.6 PSEUDO-CODE

Algorithm 1 Cognitive Gridworld Task

Input: Total latent variables $|\mathcal{S}|$, Context size C, Realizations R, Time Horizon T
Hyperparameters: Temperature λ
Initialize Embeddings: $K, Q \leftarrow$ CREATE COMPOSITIONAL EMBEDDINGS($|\mathcal{S}|$)

Create Compositional Embeddings:
$K, Q \sim \mathcal{N}(0, I)$ {Draw matrix of size $|\mathcal{S}| \times d_o \times d_\mathcal{E}$}
for $s \in \{1, \ldots, |\mathcal{S}|\}$ **do**
  $K_s \leftarrow$ GRAMSCHMIDT(K_s) {Orthogonalize across d_o}
  $Q_s \leftarrow$ GRAMSCHMIDT(Q_s)
end for
$K \leftarrow K/||K||_2, \quad Q \leftarrow Q/||Q||_2$ {Normalize across $d_\mathcal{E}$}
return K, Q

Compression & Expansion(K,Q):
$L \leftarrow 2R + 1$ {Profile length}
Initialize sinusoidal base $\mathrm{base}(n)$ of one period over L slots, scaled by λ
for ordered pair $c, c' \in \mathcal{C}$, $c \neq c'$ (or $c = c'$ if $C = 1$) **do**
  $z \leftarrow \langle \mathbf{k}^i_c, \mathbf{q}^i_{c'} \rangle$ {Get cosine similarity}
  $d(n) \leftarrow$ circular distance from n to Lz on L slots
  $\omega \leftarrow \mathrm{softmax}_n(-d(n)^2)$ {Smoothing kernel centered at Lz}
  $v^i_{cc'}(r) \leftarrow \sum_n \omega(n)\,\mathrm{base}((n - r) \bmod L)$ {Smooth roll of base}
end for
if $C = 1$ **then**
  $\ell_{\mathbf{z}^i} \leftarrow \sigma(v^i_{cc})$
else
  $\ell_{\mathbf{z}^i} \leftarrow \sigma\left(\sum_{c<c'} \text{Broadcast}(\{v^i_{cc'}(v^i_{c'c})^\top\})\right)$ {Tensor of shape R^C parameterized by Z}
end if
return $\ell_{\mathbf{z}^i}$ {Likelihood of observing $'1'$ in dimension i of $\mathbf{o}$}

Episode Generation:
Sample context indices $\mathcal{C} \subset \{1, \ldots, |\mathcal{S}|\}$, $|\mathcal{C}| = C$
Sample realizations $r_c \sim \text{Uniform}(R)$ for $c \in \mathcal{C}$
Sample goal $g \sim \text{Uniform}(\mathcal{C})$
$a^* \leftarrow r_g$
for observation dimension $i \in \mathcal{O}$ **do**
  $\ell_{\mathbf{z}^i} \leftarrow$ Compression & Expansion(K^i, Q^i)
end for
Generate trajectory $\mathbf{o}_{1:T}$ where $o^i_t \sim \text{Bernoulli}(\ell_{\mathbf{z}^i}(\mathbf{r}))$

Algorithm 2 Experiment 1: Training to capture Factorization Regret

Input: Episodes E, Interactions Z, FULLYTRAINED or ECHOSTATE Network Φ

for each episode $(\mathbf{o}_{1:T}, r_g, Z) \in E$ **do**
  $h_0 \leftarrow h_\phi$ {Initialize memory streams}
  $\mathcal{L} \leftarrow 0$ {Initialize loss}
  $\ln B_0^{\text{Net}} \leftarrow \mathbf{0}$ {Initialize log belief}
  $\ln B_0^{\text{Joint}} \leftarrow \mathbf{0}$ {Initialize log belief}
  for $t = 1$ **to** T **do**
    $[h_t, \Delta] \leftarrow \Phi([\mathbf{o}_t, Z], h_{t-1})$
    $\ln B_t^{\text{Net}} \leftarrow \ln B_{t-1}^{\text{Net}} + \Delta$
    $\ln B_t^{\text{Joint}}(\mathbf{r}) \leftarrow \ln B_{t-1}^{\text{Joint}}(\mathbf{r}) + \ln P_Z(\mathbf{o}_t|\mathbf{r}) - \ln P_Z(\mathbf{o}_t)$ {For all $\mathbf{r}$}
    $B_{tg}^{\text{Joint}} \leftarrow \sum_{\mathbf{r}_{\setminus g}} \text{Softmax}(\ln B_t^{\text{Joint}})$ {Goal marginal}
    $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{T}\mathcal{L}_{\text{FR}}$ {Using B_t, see subsection A.3}
  end for
  Update Φ, h_ϕ to minimize $\mathcal{L}$
end for

Algorithm 3 Experiment 2: Training Representation Classification Chains

Input: Episodes E, Classifier Φ, Generator Ψ
Hyperparameters: Regularization weight β
Learnable Embeddings: $\hat{K}, \hat{Q}, f_K, f_Q$

for each episode $(\mathbf{o}_{1:T}, r_g, \mathcal{C}) \in E$ **do**
  $\hat{Z} \leftarrow \langle f_K(\hat{K}_\mathcal{C}), f_Q(\hat{Q}_\mathcal{C}) \rangle$ {Estimated Interactions}
  $\ln B_0^{\text{Net}} \leftarrow \mathbf{0}$
  $h_0 \leftarrow h_\phi$

  Φ **Forward Pass:**
  for $t = 1$ **to** T **do**
    $[h_t, \Delta] \leftarrow \Phi([\mathbf{o}_t, \hat{Z}_{\text{Detached}}], h_{t-1})$
    $\ln B_t^{\text{Net}} \leftarrow \ln B_{t-1}^{\text{Net}} + \Delta$
  end for
  Sample action $a \sim B_{Tg}^{\text{Net}}$ {Action taken at $t = T$}
  $\delta \leftarrow \mathbb{I}(a = r_g)$ {Sparse reward signal}

  Φ **Backward Pass:**
  $\mathcal{L}_{Clss} \leftarrow 0$
  for $t = 1$ **to** T **do**
    $B_{tga}^{\text{Net}} \leftarrow \text{soft clip}(B_{tga}^{\text{Net}}, \epsilon_{clip})$ {Soft clip}
    $\text{Reg} \leftarrow -B_{tga}^{\text{Net}} \ln B_{tga}^{\text{Net}}$ {Regularization}
    $\text{PG} \leftarrow \delta \ln B_{tga}^{\text{Net}} + (1 - \delta) \ln(1 - B_{tga}^{\text{Net}})$ {Policy Gradient}
    $\mathcal{L}_{Clss} \leftarrow \mathcal{L}_{Clss} - \frac{1}{T}(\text{PG} + \beta \cdot \text{Reg})$
  end for
  Update Φ, h_ϕ to minimize $\mathcal{L}_{Clss}$ {Gradient flows through Φ}

  Ψ **Forward Pass:**
  $\hat{r}_g \leftarrow a$
  $\hat{r}_c \sim B_{Tc}^{\text{Net}}$ {Sample context realization}
  $\hat{\mathbf{r}}_{conf} \leftarrow B_{T\hat{\mathbf{r}}}^{\text{Net}}$, with index g set to δ {Classifier confidence}
  $\hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}}) \leftarrow \Psi([\hat{Z}, \hat{r}_g, \hat{r}_c, \hat{\mathbf{r}}_{conf}])$

  Ψ **Backward Pass:**
  $\mathbf{OPE}^i \leftarrow \frac{1}{2} J(\hat{P}_{\hat{Z}}(\mathbf{o} \mid \hat{\mathbf{r}}) \,\|\, \tilde{P}(\mathbf{o}))$ {Observation Prediction Error (soft-clipped probabilities)}
  $\mathbf{REG}_V^i \leftarrow \langle \text{MSE}(\|f_V(V^i)\|_2, 1) \rangle_c$ {Embedding Regularization}
  $\mathcal{L}_{Gen} \leftarrow \mathcal{L}_{\text{Generator}}^\star$ {Reward-weighted $\mathbf{OPE}$ plus $\mathbf{REG}$, see subsection A.4}
  Update $\Psi, \hat{K}_\mathcal{C}, \hat{Q}_\mathcal{C}, f_K, f_Q$ to minimize $\mathcal{L}_{Gen}$ {Gradient flows through Ψ to Embeddings}
end for

Algorithm 4 Experiment 3: Offline Controller Training

Input: Generator Ψ, Preferred Observation Ω (fixed for the run), a batch of contexts with estimated interactions $\{\hat{Z}\}$ (the same batch in every episode)
Hyperparameters: Training Episodes N_{ep}, Entropy bonus weight β_{cntrl}
Models: Actor Π, Critic V

Intrinsic Reward Function $\mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell)$**:**
$\ell_{\hat{\mathbf{z}}^i} \leftarrow \Psi(\hat{\mathbf{z}}^i, \mathbf{r})$ {Generator provides likelihood estimates}
$\text{PL}^i = \ln(\ell_{\hat{\mathbf{z}}^i})\Omega^i + \ln(1 - \ell_{\hat{\mathbf{z}}^i})(1 - \Omega^i)$ {Preference logit}
Reward $\leftarrow \exp\langle \text{PL} \rangle_i$
return Reward

for $e = 1$ to N_{ep} **do**
  for each context $\hat{Z}$ in the batch **do**
    PE $\leftarrow -\sum_{\mathbf{r}} \pi(\mathbf{r} \mid \hat{Z}) \ln \pi(\mathbf{r} \mid \hat{Z})$ {Compute policy entropy}
    $\mathbf{r} \sim \pi(\mathbf{r} \mid \hat{Z})$ {Sample joint action}
    $A \leftarrow \mathcal{R}(\mathbf{r}, \mathbf{\Omega}, \ell) - \nu(\hat{Z})$ {Compute Advantage}
    $\mathcal{L}_{A2C} \leftarrow A^2 - A_{\text{Detached}} \ln \pi(\mathbf{r} \mid \hat{Z}) - \beta_{cntrl} \cdot \mathbf{PE}$
  end for
  Update Π, V to minimize $\mathcal{L}_{A2C}$ averaged over the batch
end for

STATEMENT ON LLM USAGE

Large language models (LLMs) were used to assist with improving the clarity and style of the manuscript text including this statement; additionally, LLMs aided in providing relevant articles. All conceptual contributions, experimental design, and analysis were conducted by the authors.