Neural Fields as World Models

Joshua Nunley (joshnunl@iu.edu)[1,2]
[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington
[2]Cognitive Science Program, Indiana University, Bloomington

Abstract

Humans rehearse possible futures offline, as in mental practice and perhaps dreaming, suggesting that world models may support task learning away from the environment. Standard machine learning world models compress visual input into latent vectors, discarding the spatial structure that characterizes sensory cortex. We propose isomorphic world models: architectures that preserve sensory topology, so physics prediction becomes geometric propagation rather than abstract state transition. We implement this idea with motor-gated neural fields, where activity evolves through local lateral connectivity and motor commands multiplicatively modulate specific channels. Across three experiments, the same architecture learns ballistic prediction without "teleporting," improves a catching policy offline by propagating task error through a frozen learned world model, and develops body-selective motor channels without body labels. These results provide preliminary evidence that physical prediction, offline task learning, and body-linked representation share a common computational substrate: action-conditional prediction within a spatial map.

Keywords: neural fields; world models; intuitive physics; motor control; body schema

Introduction

A child learning to catch must solve two physics problems at once: where will the ball land, and where will my hand be when I reach? With practice, both predictions become automatic.

How does the brain pull this off? Battaglia, Hamrick, and Tenenbaum proposed that humans possess an "intuitive physics engine" (IPE) that runs mental simulations to predict physical outcomes (Battaglia et al., 2013). The framework accounts for human performance across diverse physical reasoning tasks (Hamrick et al., 2016; Ullman et al., 2017).

Several computational frameworks have attempted to model physical intuition. Object-based approaches like Interaction Networks (Battaglia et al., 2016) and the Neural Physics Engine (Chang et al., 2017) learn to predict physical dynamics over explicit object graphs. These models capture important structure but operate on symbolic object representations rather than raw visual input, and they do not integrate motor commands. They model physics as an observer, not as an agent. To understand how the brain predicts physics while acting in the world, we need models that integrate action. Grush (2004) proposed that the brain constructs internal emulators: forward models that predict the sensory consequences of motor commands and can run offline during imagery. Such emulators could support task improvement away from the environment, as in mental practice (Jeannerod, 1995). World models from machine learning, which learn these predictions for planning and control, can be seen as computational implementations of this idea (Ha & Schmidhuber, 2018). But these architectures omit cortical constraints. Visual input is compressed through an encoder into a latent vector with no spatial structure, a recurrent model predicts latent state transitions, and finally a decoder reconstructs predictions. In such architectures, a single weight matrix can connect any latent dimension to any other, allowing information to jump discontinuously rather than propagate locally. In the physical world, a ball moves continuously through space, forces act on adjacent objects, and trajectories unfold through spatial neighborhoods. Latent-space world models have no such constraints. Predicted objects can "teleport" across representational space from one timestep to the next.

In trajectory-prediction tasks, motion-sensitive area MT activates even without visual motion (Ahuja et al., 2022; Ahuja et al., 2024). Neural patterns during prediction resemble those during actual motion perception, as if physics simulation unfolds *within* spatially organized visual representations rather than in a separate abstract reasoning system.

The missing architecture is one that respects both constraints: spatial structure *and* action integration.

We propose that the key architectural principle is **isomorphism**: world models that preserve the spatial structure of sensory input. The defining constraint is **locality**: nearby points in the world map to nearby points in the representation, and information propagates through spatial neighbors rather than "teleporting" via global connections. When this constraint is enforced, physics prediction becomes a geometric problem: a ball's future is a path through representational space, not an abstract vector transition.

To test this hypothesis, we use neural fields, perhaps the simplest implementation of the isomorphic principle. Introduced by Amari (Amari, 1977), neural fields are spatially organized recurrent networks where activity evolves through local lateral interactions: each location influences only its spatial neighbors through a connectivity kernel. Neural fields have long served as models of cortical dynamics, capturing phenomena from visual attention to working memory to motor planning. Yet they have not been explored as world models for planning and control. Neural fields are not the only possible isomorphic world model, and 2D tasks do not exhaust physical reasoning;

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

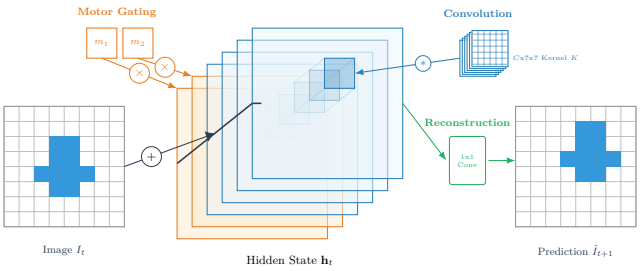

Figure 1: Neural field world model architecture. Visual input I_t is added to the hidden state only during the first three timesteps of each sequence; thereafter, the field evolves autonomously through local lateral convolution (K) and recurrence. Motor commands $\mathbf{m}$ multiplicatively gate specific channels, implementing gain modulation. A 1x1 convolution reconstructs the visual prediction $\hat{I}_{t+1}$.

here they provide a minimal, interpretable test of whether preserving spatial topology changes prediction, control, and motor-contingent representation.

To enable action-conditional prediction, we introduce *motor-gated channels*: populations whose activity is multiplicatively modulated by motor commands. The mechanism implements gain modulation, a fundamental computational principle in cortex whereby one signal multiplicatively scales neural responses to another (Salinas & Thier, 2000). In posterior parietal cortex, gain fields enable sensorimotor integration by combining visual input with motor signals (Andersen et al., 1997). Motor-gated channels implement this operation.

Together, isomorphism and motor gating define the paper's central claim. A motor-gated neural field can learn a world model that supports three cognitively relevant capacities — physical prediction, learning a new task from internal simulation, and body-linked representation — while respecting three neurobiological constraints: spatial maps, local lateral dynamics, and gain-like motor modulation.

We provide preliminary evidence for this claim in three experiments:

1. **Physics from Lateral Dynamics:** The constraint of local connectivity should make blind physical prediction proceed through continuous paths rather than discontinuous jumps (Experiment 1).

2. **Transferable Motor Control:** A frozen world model should support offline learning of a new control task: gradients pass through long coherent rollouts, and the resulting policy transfers to real physics without policy training in the real environment (Experiment 2).

3. **Body-Selective Encoding:** Motor-gated channels should develop arm-selective activity without explicit body labels, providing a minimal computational precursor to body schema (Gallagher, 2005) (Experiment 3).

Methods

Neural Field Architecture

The neural field implements the locality constraint: information propagates through spatial neighbors rather than jumping arbitrarily across the representation. At each timestep, the field state $\mathbf{h} \in \mathbb{R}^{C\times H\times W}$ updates through three influences: decay of current activity, lateral input from neighboring locations, and visual input from the environment. A 7×7 convolutional kernel K (randomly initialized, learned end-to-end with the rest of the world model) implements lateral connectivity: predicted motion must traverse intermediate positions just as objects traverse intermediate locations in physical space. The model uses an Amari-style neural field update (Amari, 1977):

$$\mathbf{h}_{t+1} = \mathbf{h}_t + \frac{\Delta t}{\tau}\left(-\mathbf{h}_t + K * \mathrm{ReLU}(\mathbf{h}_t) + W_{\text{in}} * \mathbf{I}_t\right) \quad (1)$$

Visual predictions emerge through linear reconstruction: $\hat{\mathbf{I}}_t = W_{\text{out}} * \mathbf{h}_t$.

Each channel is a complete 2D activity map over the visual field. To integrate motor commands, we designate the first M maps as motor-gated, where M matches the number of continuous motor commands ($M = 4$ for the arm; the ballistic experiment uses no motor gating). After each dynamics update, these channels are multiplicatively modulated by motor signals:

$$\mathbf{h}^{(i)}_{t+1} = m_i \cdot \tilde{\mathbf{h}}^{(i)}_{t+1} \quad \text{for } i \in \{1, \ldots, M\} \quad (2)$$

This implements gain modulation, the same computational principle by which posterior parietal cortex combines visual and motor signals (Andersen et al., 1997; Salinas & Thier, 2000).

Neurobiological constraints. The cortical comparison is structural: the model asks what follows when a world model is built from constraints common in sensory and sensorimotor cortex. It preserves three such constraints: retinotopic maps maintain spatial correspondence with visual input (Wandell et al., 2007); local lateral connectivity restricts interactions to spatial neighbors; and gain-like modulation couples motor commands to spatial activity. ReLU and convolutional kernels abstract threshold-linear activity and local connectivity. Learning is treated separately from architecture: the simulations use backpropagation, while functionally similar credit assignment may be possible in cortical circuits (Lillicrap et al., 2020).

Task Environments

Ballistic trajectory (Experiment 1). A ball moves under gravity in a 32×32 visual field, with random initial position and velocity. The model observes the first 3 frames, then predicts the remainder without visual input. Architecture: 16 channels (no motor gating).

Musculoskeletal arm (Experiments 2–3). A planar double-pendulum arm operates in a 120×45 visual field with four motor commands following the Equilibrium-Point Hypothesis (Feldman, 1986). For each joint, co-contraction (C)

activates opposing muscles together to increase stiffness; reciprocal (R) shifts the equilibrium angle to produce movement direction. The four commands are shoulder-C, shoulder-R, elbow-C, and elbow-R. These C/R variables are scalar motor commands supplied at each timestep, not labels for visual regions. A ball falls from a random horizontal position for catching. Architecture: 32 channels (4 motor-gated).

Training and Evaluation

World models are trained with mean squared prediction loss (Adam optimizer, 10,000 epochs). Frozen-simulator policy training (Experiment 2) has three stages. First, the arm world model is learned from random motor-babbling trajectories using next-frame prediction. Second, a small CNN+LSTM catch predictor is trained as a differentiable surrogate for the binary catch event. It receives visual reconstructions generated by the frozen world model while replaying real-physics rollouts; per-timestep caught/missed labels come from the corresponding real-physics trajectory. Third, the world model and catch predictor are frozen, and only the policy weights are updated. During each simulated rollout, the policy observes the reconstructed visual field, outputs motor commands, the frozen world model rolls forward, and the frozen catch predictor scores the predicted trajectory. The policy maximizes predicted catch probability through gradients that pass through both fixed modules. Thus policy learning is optimization through a learned differentiable simulator, not model-free reinforcement learning from real-environment rewards or a rollout sampler for a reward-based learner. Because gradients pass through long simulated trajectories, rollout coherence directly shapes the learned policy. The trained policy deploys to real physics without fine-tuning. Catch rate measures successful interceptions over 100 episodes.

For body schema analysis (Experiment 3), we compute a selectivity index comparing motor-gated channel activity over arm versus ball regions, normalized by reconstruction activity:

$$\text{Selectivity} = \frac{\text{Motor}_{\text{arm}}/\text{Motor}_{\text{ball}}}{\text{Recon}_{\text{arm}}/\text{Recon}_{\text{ball}}}$$

Values greater than 1 indicate body-selective representation.

Baseline. We compare against a VAE-LSTM inspired by Ha and Schmidhuber (2018) because it instantiates the architectural contrast of interest: a latent-vector world model with global recurrent dynamics, where an unstructured recurrent state can mix information from any location. A convolutional VAE compresses observations to a 32-dimensional latent space, an LSTM predicts latent dynamics, and a decoder reconstructs predictions. Following Ha and Schmidhuber (2018), training proceeds in two stages (VAE first for 10,000 epochs, then frozen VAE plus LSTM for 10,000 epochs) because gradients through the decoder destabilize joint training. This is not a parameter-matched benchmark; parameter count is descriptive, while the contrast is latent-vector versus spatially structured world-model format. The neural field requires no staged training and uses 17–67× fewer parameters (13K vs 850K for ballistic; 50K vs 1.7M for arm), largely because the VAE-LSTM requires dense layers connecting flattened features to latent space.

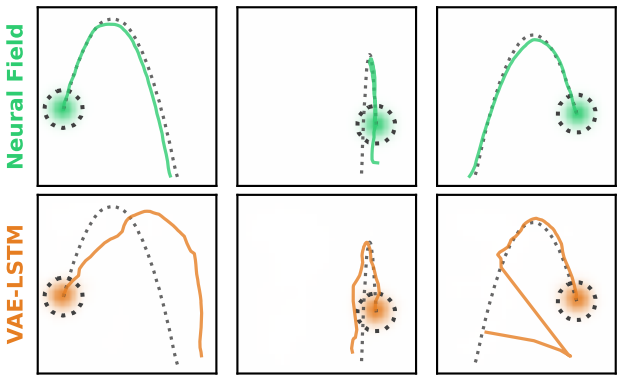

Figure 2: Trajectory prediction through internal dynamics. Each column overlays a roughly 50-timestep rollout in one plot, with successive predicted positions forming a path through time. Both models observe only the first three frames before blind prediction. The neural field (green) tracks smooth parabolic ground-truth arcs (gray dotted), while VAE-LSTM (orange) oscillates.

Results

The three experiments test increasingly action-linked consequences of preserving spatial topology. Experiment 1 asks whether local field dynamics can predict ballistic motion; Experiment 2 asks whether a frozen world model can support offline learning of a new catching policy; Experiment 3 asks whether the same motor-gated arm model contains body-selective structure without body labels.

Experiment 1: Physics from Lateral Dynamics

Figure 2 shows ballistic prediction: after observing a ball for three frames, activity in the neural field continues along the parabolic arc even without visual input. During observation, a localized bump tracks the ball; remove the input, and the bump keeps moving, following learned dynamics that approximate ballistic motion.

Across 10 seeds, the neural field achieves median prediction loss of 9.33×10^{-4} (IQR: [0.000895, 0.000988]), compared to 3.94×10^{-3} for VAE-LSTM ($p < 0.001$). Beyond aggregate loss, VAE-LSTMs exhibit a failure mode that neural fields avoid. We measured frame-to-frame displacement of predicted centroids during blind rollouts. The neural field's maximum displacement was 2.06 pixels, with 0.0% of sequences showing "teleportation" (jumps >3.0 pixels). VAE-LSTM showed 21.97 pixels maximum (10.7× larger) and 15.4% teleportation ($p < 0.001$). Local connectivity bounds how far predictions can jump in a single timestep, a constraint absent from latent-space models.

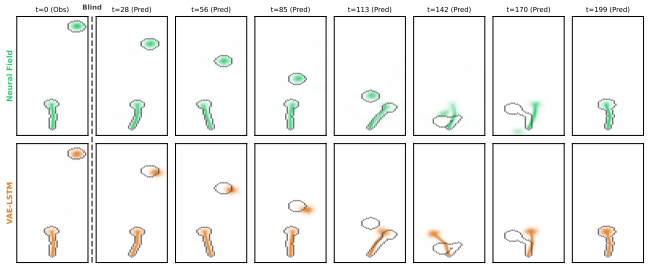

Figure 3: Visuomotor prediction in the arm catching task. Both models receive motor commands throughout, but visual input only during the first three frames (Obs). Ground truth shown as gray dots. The vertical dashed line marks the transition to blind prediction.

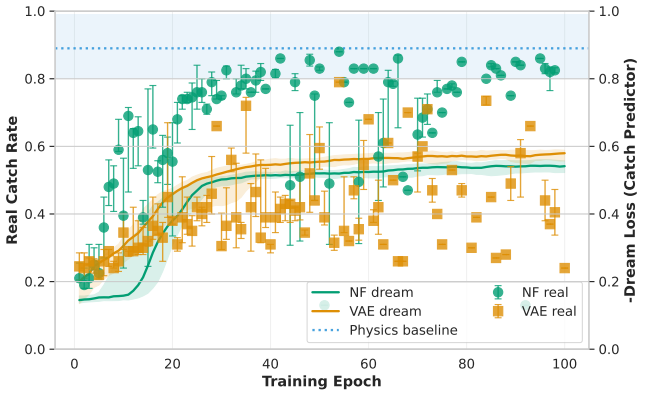

Figure 4: Frozen-simulator policy learning transfers to real physics. Lines show simulator-training catch-predictor value (right axis; negated loss, higher = better); this timestep-averaged training metric is capped near 60% because a catch can only be registered after the ball reaches catching range, and is distinct from real catch rate. Points show real catch rate at evaluation (left axis). Neural field policies (green) achieve 81.5%, approaching the physics baseline (dotted). VAE-LSTM (orange) achieves comparable simulator-training performance but shows larger sim-to-real gap.

Experiment 2: Frozen-Simulator Policy Learning

The harder test is offline task learning: can a frozen learned model support training a catching policy that works in the real environment? We freeze the trained neural field world model, then train a policy network to catch falling balls using only the model's prediction of the visual field. The policy receives the neural field reconstruction as input and outputs motor commands. Training proceeds entirely offline within the world model: the model generates predicted observations, the policy generates actions, and a fixed catch predictor (trained on real physics) scores each trajectory. Gradients pass through the frozen world model to the policy, but do not update the world model itself. This is not merely using imagined trajectories as samples; the frozen model is the differentiable simulator through which long-rollout errors shape the policy update. No real-environment interaction or additional world-model training occurs during policy learning.

Figure 3 shows the arm task: a double pendulum controlled by muscle-like actuators must catch a falling ball. The neural field receives motor commands alongside visual input during observation, then predicts the visual consequences of continued motor activity during the blind phase.

Policies trained within the neural field achieve 81.5% catch rate (IQR: [79.2%, 84.7%]) when deployed to the *real* physics environment, approaching the 89.0% achieved by policies trained directly on real physics (Figure 4). VAE-LSTM policies achieve only 46.0% ($p = 0.003$), roughly half the neural field's performance.

Prediction loss alone does not determine transfer quality: despite comparable world model losses (Figure 5), the neural field supports substantially better policy transfer. MSE measures visual prediction fit; rollout continuity and real catch rate are the behaviorally relevant tests.

Why does spatial structure help? The neural field's dynamics function as a spatial integrator: position and velocity are represented in the location and movement of activity, and the learned kernel implements discrete integration over time. In latent-space models, these quantities must be disentangled from a compressed representation. Spatial inductive bias provides a natural coordinate system for physics, allowing the policy to exploit spatial features directly rather than learning to decode them from an arbitrary latent space.

Experiment 3: Emergent Body-Selective Encoding

Body schema, the implicit representation of one's own body that enables coordinated action, has long been studied as a distinct cognitive capacity (Gallagher, 2005). In the arm model, a computational precursor to body schema emerges as a byproduct of visuomotor prediction. The motor-gated architecture supplies action-modulated channels but does not specify which pixels represent the body or where selectivity should emerge; arm-versus-ball localization is measured only after training. The finding provides computational support for developmental theories proposing that infants discover their bodies through sensorimotor contingencies (Bahrick, 1995; Rochat, 1998). Using the arm world model, we compare motor-gated channel activity over the arm versus the ball, normalized by reconstruction activity.

Reciprocal (R) motor channels, which control joint movement direction, show significant arm selectivity (Figure 6): shoulder R median selectivity = 2.18 ($p = 0.002$); elbow R median = 1.50 ($p = 0.002$). These channels are approximately 2× more active over the arm than reconstruction alone would require. Co-contraction (C) channels show no significant selectivity (shoulder C: $p = 0.16$; elbow C: $p = 0.96$).

Why do R channels show selectivity while C channels do not? The answer likely lies in visual consequences. Reciprocal commands produce large arm movements; to predict these, the motor channel must represent where the arm is.

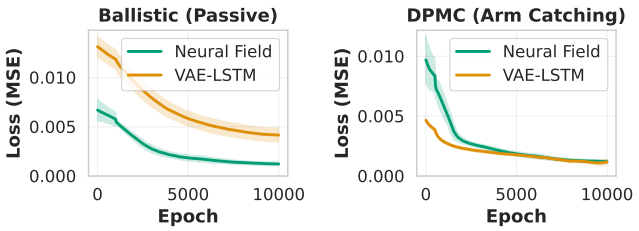

Figure 5: Training dynamics. Neural field (green) achieves lower final pixel-MSE loss than VAE-LSTM (orange) on the ballistic task and comparable loss on the arm task, despite using 17–67× fewer parameters. Task-relevant metrics are reported as centroid continuity and real catch rate. Shaded regions show IQR across seeds.

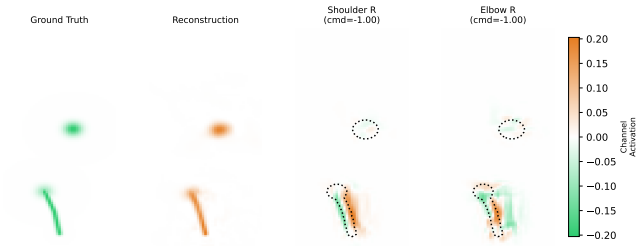

Figure 6: Emergent body-selective encoding in motor-gated channels. Reciprocal (R) channels show activation concentrated over the arm rather than the ball (dotted contours), despite no explicit training to distinguish body from world. Orange indicates positive activation, green negative.

Co-contraction modulates stiffness with minimal immediate visual change. The selectivity pattern reflects what each motor command *does*. Body selectivity emerges from the prediction objective, not explicit supervision.

Motor-gated channels show how neural fields spatialize dynamics. Each channel encodes action-contingent movement within a spatial pattern. Channel activation at any instant represents where the controlled object *will move*, pixel by pixel, under the current motor command. The same arm configuration produces different channel patterns depending on which motor is active. The architecture transforms temporal dynamics into spatial structure, representing "what will move where" in the spatial pattern of the field itself.

Discussion

We proposed that motor-gated neural fields can learn world models that support three cognitively relevant capacities while respecting three neurobiological constraints: spatial maps, local lateral dynamics, and gain-like motor modulation. Three results provide preliminary evidence. First, local connectivity proves sufficient to learn ballistic physics: trajectories emerge from wave-like propagation through the field, and predictions traverse intermediate locations rather than teleporting as in latent-space models (Experiment 1). Second, the spatial structure supports learning a new control task through a frozen differentiable simulator: policies trained through neural-field rollouts transfer to reality with a smaller sim-to-real gap than the latent baseline (Experiment 2). Third, a computational precursor to body schema emerges from visuomotor prediction alone. Motor-gated channels develop body-selective encoding without explicit body labels (Experiment 3).

Relation to prior work. The neural field offers a mechanistic substrate for the "intuitive physics engine" proposed by Battaglia et al. (Battaglia et al., 2013): physics emerges from learned connectivity rather than explicit rules. Ahuja et al. found that area MT activates during trajectory prediction even without visual motion (Ahuja et al., 2022). Our neural field shows analogous behavior: during blind prediction, activity moves through retinotopic space, providing a computational parallel to MT's motion representation. Multiplicative gating by motor signals implements gain modulation, the same computational principle found in posterior parietal cortex where motor signals modulate spatially organized sensory representations (Andersen et al., 1997; Salinas & Thier, 2000). Neural fields have a rich history in cognitive science through Dynamic Field Theory (DFT), which hand-specifies Amari-type dynamics to model working memory, motor planning, and decision-making (Schöner et al., 2016). Our work departs from this tradition: rather than designing field dynamics for specific cognitive functions, we learn them end-to-end from sensorimotor prediction. The neural field learns its own dynamics, including physics and body-selective encoding, through prediction error within this structured architecture. On this view, computational principles identified by DFT can be learned when spatially structured networks predict the sensory consequences of action.

Neural fields and VAE-LSTMs can achieve similar loss, but neural fields avoid the failure mode that affects VAE-LSTMs. Local connectivity constrains errors to smooth spatial drift. Predictions cannot jump arbitrarily across the field because information must propagate through spatial neighbors. This is not only an engineering difference: architectural constraints shape how models fail. Human multiple object tracking errors arise primarily from spatial proximity. For example, when targets approach within a few degrees, crowding causes confusions (Franconeri et al., 2010). Neural fields with local connectivity should exhibit similar proximity-dependent interference, a prediction we leave for future work.

The VAE-LSTM comparison is a contrast between representational formats, not an exhaustive benchmark against all world models: spatially structured alternatives such as ConvLSTMs may share these benefits, which would support the broader isomorphism hypothesis. The key claim is that preserving spatial topology changes how prediction errors unfold and whether a frozen model can train policies for new tasks through differentiable long rollouts.

Implications. The body-selectivity result has a simple implication: body-linked sensorimotor representation need not be explicitly labeled. One might object that this is circular:

gated channels will represent whatever moves with their gating signal. The objection assumes, however, that the architecture supports stable visuomotor prediction in the first place; neither multiplicative gating nor purely local connectivity guarantee this. The claim is therefore conditional: when motor-gated neural fields succeed at visuomotor prediction, body-selective encoding follows.

The result instantiates one form of contingency detection (Bahrick & Watson, 1985): given only action-conditional prediction, the distinction between controlled body and external object emerges from sensorimotor statistics. Infants as young as three months preferentially attend to contingent visual feedback from their own limb movements (Bahrick, 1995; Rochat, 1998). Our model offers a candidate mechanism by which this contingent attention could give rise to body-selective representations. The motor-gated architecture provides a mechanistic implementation. Channels gated by motor commands can develop representations of the body because the body is what moves contingently with those commands. This can be read as a simple comparator-like mechanism for agency (Blakemore et al., 2002), which proposes that the brain distinguishes self-caused from externally-caused sensations by comparing predicted sensory consequences against actual feedback. Motor-gated channels generate action-conditional predictions; when these match observed motion, prediction error is low, while mismatch signals externally-caused motion.

Grush's emulation theory (Grush, 2004) gives Experiment 2 its cognitive analogue: the brain constructs internal models (emulators) that can be run offline to simulate action outcomes. During overt movement, these emulators run in parallel with actual sensorimotor loops; during imagery, they run alone. Our neural field functions as such an emulator, a learned forward model that predicts sensory consequences of motor commands. The key procedural distinction is that the emulator is not updated during policy learning, nor used merely to sample rollouts for a reward-based learner; it is the fixed differentiable system through which gradients train a new policy. This imposes a stricter demand than short-term prediction: long rollouts must remain coherent because errors in the simulator's rollouts directly shape the policy gradients. The training order also has a cognitive analogue: outcome recognition can be acquired from sparse exploratory feedback before a competent policy exists, then used to refine that policy offline. Frozen-simulator learning succeeds because the emulator is accurate enough that policies trained offline transfer to reality. Mental practice (Jeannerod, 1995) is a documented biological analogue, and dream-like rehearsal may be another: offline uses of internal forward models that improve action without engaging the body. We model this computational role, not sleep or imagery itself.

The representational format matters here. Latent-space models function as descriptive systems. They compress experience into abstract codes where spatial relationships are discarded and must be reconstructed through decoding. The model must recover spatial relations that its representation has thrown away. Isomorphic world models are constitutive, meaning that representation and world share geometric form. To predict a trajectory is to propagate activity through space; to represent position is to activate a location. The neural field is therefore directly interpretable: one can watch field dynamics unfold and read off the predicted position as activity in a particular location, not as a latent vector requiring decoding. This aligns with enactivist accounts of perception (O'Regan & Noë, 2001), where understanding arises from mastery of sensorimotor contingencies rather than construction of internal pictures. Experiment 3 illustrates this point directly. The motor-gated channels learn not a symbol for "arm" but the sensorimotor contingency itself. They encode what changes when a particular action is commanded. Because the representation is spatial, the answer is spatially readable, which may explain why physical reasoning feels immediate rather than inferential.

Limitations and future directions. Our demonstrations involve simple 2D environments. Object interaction, including occlusion, collision, and support, may require representing objects in separate layers with gated interactions, so they move independently while influencing each other through learned coupling. More fundamentally, isomorphism does not need to mean pixel-isomorphism: the neural field could operate in a three-dimensional coordinate system that preserves the spatial structure of physical space rather than retinal space. There is biological precedent: hippocampal place cells represent volumetric 3D space (Grieves et al., 2020), and parietal area V6A encodes reaching targets in depth coordinates (Hadjidimitrakis et al., 2014). Visual input could project into this volumetric representation, and dynamics would unfold in a space where depth is explicit. Such an architecture would remain isomorphic by preserving spatial adjacency in the physical world while handling the projective distortions that make 2D pixel space a poor model of 3D physics.

The observed benefits–stable policy gradients, successful simulator-to-real transfer, and body-selective encoding–likely arise from spatial adjacency plus action modulation rather than neural-field dynamics alone. The broader class includes ConvLSTMs, cellular automata, and graph neural networks on spatial graphs–any architecture preserving spatial adjacency–and motor gating is one of several possible mechanisms for action integration. Benchmarking this class to identify which architectural choices matter most for physics prediction is future work.

Intuitive physics supports but does not exhaust physical reasoning. The architecture we propose handles the fast, automatic simulation supporting perception and action; symbolic reasoning about physics likely requires additional mechanisms, and how neural field world models might interface with such systems remains an open question.

Conclusion. A motor-gated neural field treats physics, control, and body representation as one spatial problem: predicting what changes where when an agent acts.

Acknowledgments

This work used the Big Red 200 supercomputer at Indiana University, supported by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. The author thanks Brandon Nunley for helpful discussions. AI tools assisted with editing the manuscript and writing portions of the code; the author reviewed all AI-generated content, takes full responsibility for the final work, and validated the code through systematic testing and visualization.

References

Ahuja, A., Desrochers, T. M., & Sheinberg, D. L. (2022). A role for visual areas in physics simulations. *Cognitive Neuropsychology*, *38*(7-8), 425–439. https://doi.org/10.1080/02643294.2022.2034609

Ahuja, A., Rodriguez, N. Y., Ashok, A. K., Serre, T., Desrochers, T. M., & Sheinberg, D. L. (2024). Monkeys engage in visual simulation to solve complex problems. *Current Biology*, *34*(24), 5635–5645. https://doi.org/10.1016/j.cub.2024.10.026

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, *27*(2), 77–87.

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*(1), 303–330.

Bahrick, L. E. (1995). Intermodal origins of self-perception. In P. Rochat (Ed.), *The self in infancy: Theory and research* (pp. 349–373). North-Holland/Elsevier.

Bahrick, L. E., & Watson, J. S. (1985). Detection of intermodal proprioceptive–visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, *21*(6), 963–973.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D., & Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, *29*, 4502–4510.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in cognitive sciences*, *6*(6), 237–242.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2017). A compositional object-based approach to learning physical dynamics. *International Conference on Learning Representations*.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of Motor Behavior*, *18*(1), 17–54.

Franconeri, S. L., Jonathan, S. V., & Scimeca, J. M. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*(7), 920–925.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford University Press.

Grieves, R. M., Jedidi-Ayoub, S., Mishchanchuk, K., Liu, A., Renaudineau, S., & Jeffery, K. J. (2020). The place-cell representation of volumetric space in rats. *Nature Communications*, *11*(1), 789. https://doi.org/10.1038/s41467-020-14611-7

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396.

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, *31*, 2451–2463.

Hadjidimitrakis, K., Berber, M. F., Sheth, A., Bhansali, P., Breveglieri, R., Bosco, A., Fattori, P., & Galletti, C. (2014). Common neural substrate for processing depth and direction signals for reaching in the monkey medial posterior parietal cortex. *Cerebral Cortex*, *24*(6), 1645–1657.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia*, *33*(11), 1419–1432.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Rochat, P. (1998). Self-perception and action in infancy. *Experimental Brain Research*, *123*(1), 102–109.

Salinas, E., & Thier, P. (2000). Gain modulation: A major computational principle of the central nervous system. *Neuron*, *27*(1), 15–21.

Schöner, G., Spencer, J., Group, D. R., et al. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383.