# What Do World Models Learn in RL? Probing Latent Representations in Learned Environment Simulators

Xinyu Zhang
Anyscale
xinyu@gmail.com

## Abstract

World models learn to simulate environment dynamics from experience, enabling sample-efficient reinforcement learning. But what do these models actually represent internally? We apply interpretability techniques—including linear and nonlinear probing, causal interventions, and attention analysis—to two architecturally distinct world models: IRIS (discrete token transformer) and DIAMOND (continuous diffusion UNet), trained on Atari Breakout and Pong. Using linear probes, we find that both models develop linearly decodable representations of game state variables (object positions, scores), with MLP probes yielding only marginally higher $R^2$, confirming that these representations are approximately linear. Causal interventions—shifting hidden states along probe-derived directions—produce correlated changes in model predictions, providing evidence that representations are functionally used rather than merely correlated. Analysis of IRIS attention heads reveals spatial specialization: specific heads attend preferentially to tokens overlapping with game objects. Multi-baseline token ablation experiments consistently identify object-containing tokens as disproportionately important. Our findings provide interpretability evidence that learned world models develop structured, approximately linear internal representations of environment state across two games and two architectures.

## 1 Introduction

World models—learned simulators of environment dynamics—have become a cornerstone of sample-efficient reinforcement learning (Ha & Schmidhuber, 2018; Hafner et al., 2023). Recent advances such as IRIS (Micheli et al., 2023) and DIAMOND (Alonso et al., 2024) achieve strong performance on the Atari 100k benchmark by learning to predict future observations entirely from experience. Yet despite their empirical success, a fundamental question remains: what do world models represent internally?

This question connects to the broader "linear representation hypothesis" emerging in mechanistic interpretability research (Park et al., 2023; Elhage et al., 2022). Li et al. (2023) showed that a transformer trained to predict Othello moves develops an emergent linear representation of the board state, despite never being trained on it directly. Nanda et al. (2023) extended this to chess. If sequence models trained on game transcripts develop world representations, do world models—which are explicitly trained to predict future observations—develop even richer ones?

We apply probing classifiers (Alain & Bengio, 2017; Belinkov, 2022) and causal interventions to two architecturally distinct world models—IRIS (Micheli et al., 2023) (VQ-VAE + transformer) and DIAMOND (Alonso et al., 2024) (UNet diffusion)—on Breakout and Pong. We contribute: (1) linear and MLP probing showing approximately linear game state representations ($\Delta \leq 0.06$); (2) causal interventions confirming representations are functionally used ($r > 0.95$); (3) per-head spatial attention specialization; and (4) multi-baseline token ablation with consistent importance rankings ($\rho > 0.9$).

# 2 Method

## 2.1 Models and Ground Truth

IRIS tokenizes $64 \times 64$ observations into 16 discrete tokens (VQ-VAE, codebook 512, $4 \times 4$ grid) then predicts sequences with a GPT-2 transformer (10 layers, 4 heads, dim 256). DIAMOND uses a UNet denoiser (4 stages, 64 channels) with EDM preconditioning. We probe on Breakout (ball_x, ball_y, player_x, score) and Pong (ball_x, ball_y, player_y, enemy_y), with ground truth from Atari RAM (Anand et al., 2019).

## 2.2 Probing Protocol

We extract frozen representations from all layers ($N$=10,000 frames per game): IRIS VQ-VAE encoder/embedding + 10 transformer layers; DIAMOND conv input, 4 encoder/decoder stages, bottleneck, norm output. For each (layer, property) pair, we train Ridge regression ($\alpha$=1.0) and 2-layer MLP probes (256→128→1, ReLU, Adam), both with 5-fold CV $R^2$. The selectivity gap $\Delta = R^2_{\text{MLP}} - R^2_{\text{linear}}$ measures nonlinear structure. Controls: raw pixels, random model, shuffled labels.

## 2.3 Causal Intervention Protocol

To move beyond correlation, we perform activation patching along probe directions (Geiger et al., 2021): modify hidden states $\mathbf{h}' = \mathbf{h} + \alpha \cdot \hat{\mathbf{w}}$ (where $\hat{\mathbf{w}}$ is the normalized Ridge probe weight) and measure the resulting change in next-token logits (KL divergence, token change rate). A positive correlation between $|\alpha|$ and prediction change indicates the probe direction is functionally used.

## 2.4 Attention Analysis and Token Ablation

For IRIS's 40 attention heads, we compute attention entropy and per-head spatial selectivity (mean attention to each of 16 token positions, forming $4 \times 4$ maps). For token ablation, we replace each spatial token under three baselines (zero, mean, random codebook entry) and measure prediction disruption (KL divergence). Consistency across baselines (Spearman $\rho$) indicates robust importance rankings.

# 3 Results

## 3.1 Linear Representations Across Games

Figure 1 shows probe $R^2$ across all layers for both games.

IRIS. Ball position encoding is stable across all transformer layers in both games: for Breakout, $R^2$ for ball_x ranges from $0.84 \pm 0.01$ (layer 0) to $0.85 \pm 0.006$ (layer 5), a span of only 0.01. Paddle position and score are near-perfectly encoded ($R^2 > 0.99$) at every layer. The VQ-VAE encoder already achieves $R^2 = 0.83$ for ball-x, and the transformer barely improves this (+0.02).

DIAMOND. Early encoder stages (0–2) show negative $R^2$ for ball position (as low as $-1.45$), while the deepest encoder stage begins recovering ($R^2 = 0.78$); the bottleneck achieves peak linear $R^2 = 0.81 \pm 0.01$ for ball_x, and decoder stages decline—an inverted-V pattern suggesting the bottleneck compresses information maximally. MLP probes recover substantially more from decoder layers ($R^2 = 0.91$ at dec_2 vs. $-0.27$ linear), indicating these layers encode ball position nonlinearly via skip connections.

Cross-game consistency. Pong shows the same architectural signatures: flat IRIS profiles and peaked DIAMOND bottleneck (Figure 1, bottom row), indicating these patterns are architecture-dependent rather than game-specific. However, both models track ball position significantly better in Pong ($R^2 \geq 0.95$) than Breakout ($R^2 \leq 0.85$), likely because Pong's

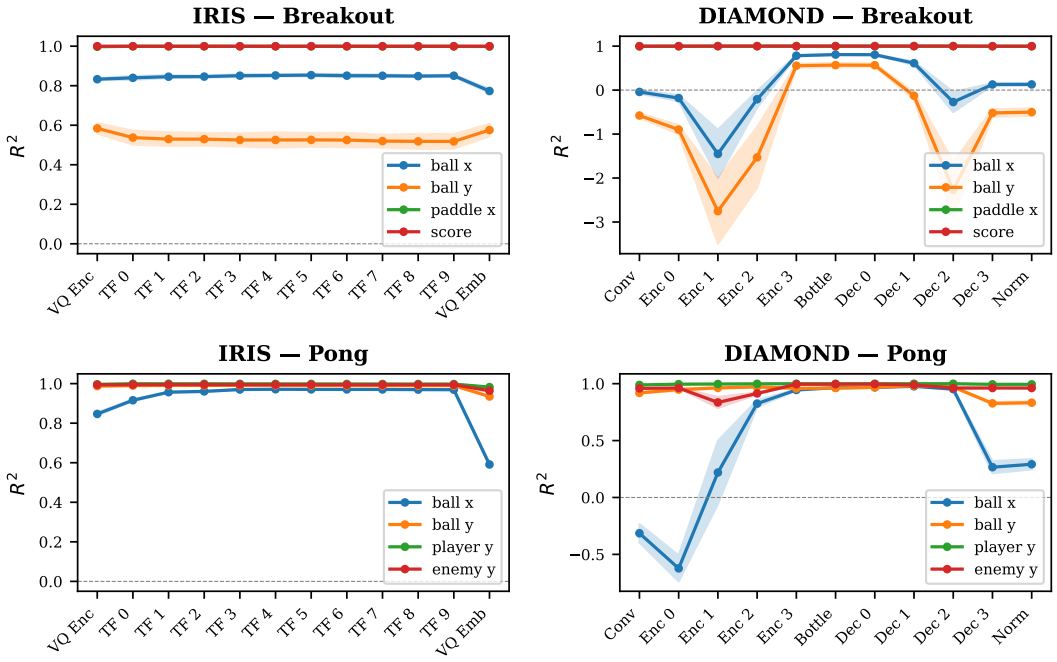


Figure 1: Probe $R^2$ across layers (in network data-flow order) for IRIS (left) and DIAMOND (right) on Breakout (top) and Pong (bottom). Each line tracks one game-state property; shaded bands show $\pm 1$ std over 5-fold CV. IRIS representations are flat across transformer layers, while DIAMOND shows a peaked inverted-V centered on the UNet bottleneck. Note: $y$-axis includes negative $R^2$ values, revealing that DIAMOND's early encoder layers are worse than a constant predictor for ball position.

Table 1: Best-layer $R^2$ (mean $\pm$ std over 5-fold CV) for Breakout. Both linear and MLP probes are shown; the small selectivity gap ($\Delta$) confirms approximately linear representations.

| Representation | ball_x | ball_y | player_x | score |
|---|---|---|---|---|
| Random model | $-1.21$ | $-1.22$ | $-1.14$ | $-1.18$ |
| Shuffled labels | $-0.51$ | $-0.49$ | $-0.53$ | $-0.52$ |
| Raw pixels | $-1.31$ | $-0.48$ | $0.9989 \pm .0006$ | $0.9998 \pm .0001$ |
| IRIS (Linear) | $\mathbf{0.85} \pm .006$ | $0.58 \pm .03$ | $\mathbf{0.9994} \pm .0001$ | $\mathbf{0.9999} \pm .0001$ |
| IRIS (MLP) | $0.91 \pm .005$ | $0.59 \pm .03$ | $0.9987 \pm .0003$ | $0.9999 \pm .0000$ |
| $\Delta_{\text{IRIS}}$ | $+0.06$ | $+0.01$ | $-0.0007$ | $+0.0000$ |
| DIAMOND (Linear) | $0.81 \pm .01$ | $\mathbf{0.57} \pm .05$ | $1.0000 \pm .0000$ | $0.9999 \pm .0000$ |
| DIAMOND (MLP) | $0.91 \pm .005$ | $0.63 \pm .05$ | $0.9994 \pm .0002$ | $0.9998 \pm .0001$ |
| $\Delta_{\text{DIAMOND}}$ | $+0.10$ | $+0.06$ | $-0.0006$ | $-0.0001$ |

visual scene is simpler (no bricks, fewer objects). Interestingly, DIAMOND's V-shape in ball tracking is much less pronounced in Pong (dec_2 $R^2 = 0.95$ vs. $-0.27$ in Breakout), suggesting this pattern is game-dependent.

Table 1 shows that MLP probes yield only marginal improvements over linear probes for IRIS ($|\Delta| \leq 0.06$). DIAMOND shows a larger gap for ball position ($\Delta = 0.10$ for ball_x), driven by decoder layers where skip connections mix information nonlinearly; at the bottleneck itself, the gap is small ($\Delta = 0.04$). Both models dramatically outperform baselines, with raw pixels failing on ball position ($R^2 = -1.31$), showing that world model representations extract non-trivial structure.

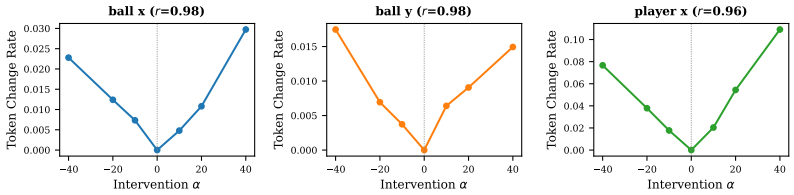


Figure 2: Causal intervention on Breakout: shifting IRIS layer-5 hidden states along probe directions produces correlated changes in predictions ($r \geq 0.96$ for all properties, measured via KL divergence).

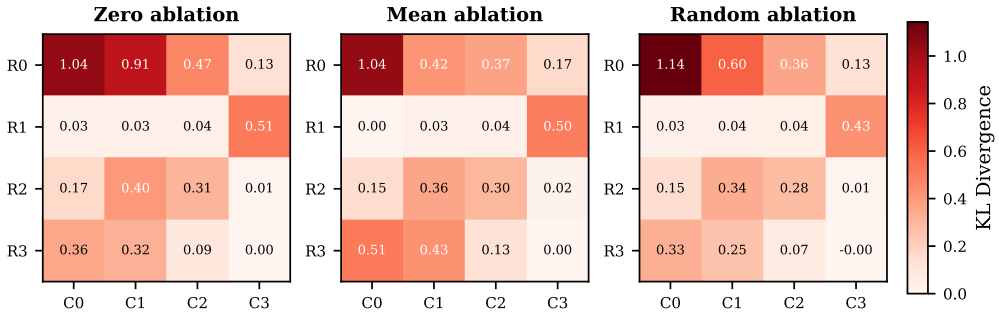


Figure 3: Three-way token ablation on Breakout ($4 \times 4$ grid). Zero, mean, and random replacement produce consistent importance rankings ($\rho > 0.92$), with token 0 (score/brick region) most critical.

For Pong, both models achieve even higher $R^2$: IRIS tracks ball position with $R^2 = 0.97$ (linear) and 0.995 (MLP); DIAMOND achieves $R^2 = 0.98$ (linear) and 0.99 (MLP). Selectivity gaps are uniformly small ($\Delta \leq 0.03$), confirming approximately linear encoding across both games.

### 3.2 Causal Interventions Confirm Functional Use

Figure 2 shows that shifting IRIS layer-5 hidden states along probe directions produces monotonically increasing prediction changes. Correlation between $|\alpha|$ and KL divergence is strong: $r = 0.97$ (ball_x), $r = 0.97$ (ball_y), $r = 0.97$ (player_x). player_x interventions produce $\sim$16$\times$ larger KL than ball interventions (0.033 vs. 0.002 at $\alpha$=40), suggesting the model relies more on paddle position. This confirms that linear representations are functionally used, not mere artifacts.

### 3.3 Attention and Token Ablation

Attention entropy ranges from 1.0–1.75 nats across 40 heads (below $H_{\max} = 2.83$), and individual heads show distinct spatial preferences: the four most selective heads—$(0, 3)$, $(4, 2)$, $(6, 0)$, $(5, 0)$—concentrate attention on different spatial regions, suggesting division of labor for tracking game elements.

Token ablation (Figure 3) consistently identifies token 0 (score/brick region) as most critical (KL $> 1.0$, $\sim$50% token change rate). Rank correlation across methods is high ($\rho = 0.93$ zero/mean, $\rho > 0.99$ zero/random). KL divergence correlates moderately with ball distance ($r \approx 0.56$ for Breakout), while Pong shows weaker spatial correlation ($r \approx 0.13$), suggesting information is distributed less spatially in simpler scenes.

## 4 Discussion and Conclusion

Our results demonstrate that learned world models develop structured, approximately linear internal representations of game state across two games and two architectures. This parallels

findings from the Othello-GPT line of work (Li et al., 2023; Nanda et al., 2023), extending them to pixel-based environment simulation.

Architectural comparison. IRIS's VQ-VAE tokenizer already produces strong linear representations ($R^2 = 0.83$ for ball position), which the transformer preserves but barely improves ($R^2 = 0.85$, a gain of only $+0.02$). Rather than a limitation, this reveals a meaningful division of labor: the tokenizer handles spatial encoding while the transformer focuses on temporal dynamics and prediction—a factorization that single-frame probes cannot fully evaluate. DIAMOND concentrates abstract state sharply at the UNet bottleneck; we hypothesize skip connections allow low-level information to bypass it, so only the bottleneck must encode abstract state.

Limitations. Both games are 2D Atari; generalization to 3D environments remains open. Single-frame probes may miss temporal structure that the transformer encodes; sequence-conditioned probes could reveal richer dynamics. Activation patching along a single direction is a coarse intervention; more targeted causal methods (Geiger et al., 2021) could strengthen these findings.

## References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In International Conference on Learning Representations (Workshop), 2017.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storber, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In Advances in Neural Information Processing Systems, 2024.

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. In Advances in Neural Information Processing Systems, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1):207–219, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. arXiv preprint arXiv:2209.10652, 2022.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In Advances in Neural Information Processing Systems, 2021.

David Ha and Jürgen Schmidhuber. World models. In Advances in Neural Information Processing Systems, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. In International Conference on Machine Learning, 2023.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In International Conference on Learning Representations, 2023.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world learners. In International Conference on Learning Representations, 2023.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. arXiv preprint arXiv:2309.00941, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. arXiv preprint arXiv:2311.03658, 2023.