# Cross-Game Semantic Alignment of Latent Action Representations in World Models

Olufeolu Kolawole
Human Perception Lab, Stanford University

Khizer Khaderi
Human Perception Lab, Stanford University

March 2026

## Abstract

World models learn latent action representations from unlabeled video, but it is unclear whether these representations capture semantically consistent actions across different games. We investigate this by training a single inverse dynamics model (IDM) jointly across multiple games and examining the geometry of the resulting embedding space. Across three domains — racing games, Atari games, and first-person games — we find a consistent pattern: the model aligns actions that produce similar visual effects (e.g., "steer left" clusters together regardless of game), but alignment degrades when the same action looks different across games. We call this degradation a *calibration gap*: the model encodes the right information, but the boundary between action classes shifts from game to game. On racing games, the joint IDM reaches 80.2% on Pole Position and 58.3% on Enduro with a single set of weights; the remaining gap to the 73.7% supervised ceiling is entirely a calibration problem. On first-person games, movement keys align strongly (85.2% Minecraft, 77.8% CS:GO) while camera direction shows a larger calibration gap that is fully recoverable with 50 labeled frames. We show that this calibration gap can be closed without any labels by estimating thresholds from optical flow statistics alone, recovering 91.1% accuracy on Pole Position.

## 1 Introduction

Unsupervised world models [4, 8, 1] learn to predict what will happen next in a video by discovering latent "actions" — internal representations of what caused the scene to change from one frame to the next. A natural question follows: do these learned actions actually mean anything? If a model watches both Pole Position and Enduro gameplay, does it learn that "steer left" is the same action in both games, or does it memorize two unrelated patterns?

The answer matters practically. If actions align across games, then a model trained on game A can transfer its understanding to game B without any new labels. If they don't, every game requires its own model from scratch.

This paper studies when cross-game alignment emerges and when it breaks down. We train a single IDM on multiple games simultaneously and examine the geometry of its embedding space using t-SNE, linear probes, and cross-game KNN accuracy. Across three domains of increasing difficulty — racing games, Atari games, and first-person games — we find the same pattern:

- Actions that produce consistent visual signatures (e.g., road edges shifting for steering) align reliably across games.
- When alignment partially fails, it is not because the model missed the information — it is because the *boundary* between action classes sits at different thresholds in different games. We call this a **calibration gap**.
- The calibration gap is closable: 50 labeled frames per class recover full accuracy, and in some cases, flow statistics alone close it with zero labels.

**Contributions.**

- A systematic study of cross-game action alignment across three domains, showing that a single IDM clusters embeddings by action class rather than game identity.
- Identification of the calibration gap as the primary bottleneck: the model's representation is sound, but per-game thresholds are needed at the decision boundary.
- A cross-game KNN similarity metric that predicts which game pairs will benefit from joint training before running the full experiment.
- A zero-label calibration method using flow statistics that recovers 91.1% accuracy on Pole Position (within 0.3 pp of the supervised ceiling).

## 2 Related Work

**Latent action world models.** Genie [4] learns latent actions from unlabeled video via a learned action tokenizer. LAPO [11] jointly trains an IDM and a forward

dynamics model so that latent actions must be both predictable from frame pairs and useful for generation. AdaWorld [7] extends this to adaptable world models that transfer to novel environments. These works focus on generation quality; we focus on whether the learned action space is semantically consistent across games.

**Inverse dynamics for representation learning.** Brandfonbrener et al. [3] showed that IDM pretraining outperforms other self-supervised objectives for learning transferable representations. VPT [2] trains a supervised IDM on Minecraft contractor data and achieves strong single-game results. We extend the IDM to the cross-game setting and study the geometry of the shared embedding space.

**Cross-game behavior alignment.** BehAVE [10] aligns video encodings across 25 first-person games using textual behavior descriptions. We take a different approach: rather than aligning with text, we train a joint IDM and diagnose *why* alignment succeeds or fails by analyzing embedding geometry directly. Domain-adversarial training (DANN [6]) is the standard approach for cross-domain feature alignment; we find that joint IDM training achieves comparable results without adversarial instability.

# 3 Method

All experiments use the same architecture and evaluation framework, applied to progressively harder game domains.

**Model.** We use IDMLatentSpatial, a 4-layer CNN with InstanceNorm2d that operates on 16×16 spatial feature grids. The model takes a pair of frames as input and predicts the action that caused the transition. All games share a single set of weights — the model receives no game identity signal.

**Training.** For racing and first-person games, the IDM is trained on raw frame pairs with ground truth player input labels recorded during human gameplay. For Atari games, where emulators do not log raw input events, we compute Farnebäck optical flow [5] between consecutive frames and assign 3-class direction labels (LEFT / NEUTRAL / RIGHT) based on the dominant flow angle. Horizontal flip augmentation (p=0.5) is used throughout to mitigate class imbalance.

**Evaluation.** We assess alignment with three complementary tools:

- **t-SNE visualization**: do embeddings cluster by action class or by game identity?
- **Linear probes**: can a simple classifier recover action labels from the IDM's penultimate layer? This tests whether information is *present* in the embedding, independent of the model's own decision boundary.
- **Cross-game KNN accuracy**: for a frame from game A, do its nearest neighbors from game B share the same action label? This directly measures cross-game alignment.

# 4 Experiments

We apply this framework to three domains of increasing difficulty. In each domain, we observe the same pattern: strong alignment for actions with consistent visual signatures, and a calibration gap for actions whose visual appearance varies across games.

## 4.1 Racing Games

**Setup.** We collected 33,770 labeled frames of Pole Position and 2,453 labeled frames of Enduro, each with ground truth key labels [LEFT, RIGHT, ACCEL, BRAKE] recorded directly from player input. Pole Position features strong edge-based steering flow (road boundaries shift visibly with each turn), while Enduro has more complex scroll and parallax effects.

**Cross-game alignment.** The joint model achieves **80.2%** balanced accuracy on Pole Position and **58.3%** on Enduro (Table 1) using a single set of weights. Figure 1 shows t-SNE embeddings forming three clean clusters (LEFT, NEUTRAL, RIGHT) with PP and Enduro frames *interleaved within each cluster* — the model has learned to represent steering direction, not game identity.

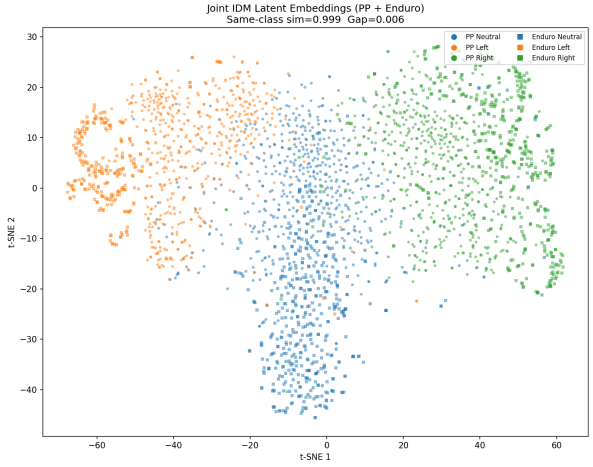


Figure 1: t-SNE of joint IDM embeddings for Pole Position (circles) and Enduro (triangles), colored by steering direction. Frames from both games are interleaved within each action cluster, confirming semantic alignment.

Table 1: Balanced accuracy on racing games.

| Model | PP | Enduro |
|---|---|---|
| Chance | 33.3% | 33.3% |
| Single-game (PP only) | 88.1% | — |
| Single-game (Enduro only) | — | 52.1% |
| Joint IDM v1 | 77.2% | 51.4% |
| Joint IDM v2 (+balanced sampling) | 76.6% | 59.3% |
| **Joint IDM v3 (+hflip aug.)** | **80.2%** | **58.3%** |
| Calibrated ceiling (20% labels) | — | 73.7% |

**The calibration gap (first appearance).** The joint IDM reaches 58.3% on Enduro, but fine-tuning with just 20% of Enduro's labels pushes this to 73.7% — a 15.4 pp gap. Crucially, the t-SNE clusters are already well-separated; the gap is not in the representation but in the decision boundary. Enduro has 10× fewer left-turn examples than Pole Position, so the model's learned threshold for "what counts as neutral" is biased toward PP's distribution. This is the first instance of the calibration gap — the same phenomenon reappears in first-person games at larger scale.

## 4.2 Atari Games

**Setup.** We use the Atari-HEAD dataset [12], which provides human gameplay for 17 Atari games. As noted in Section 3, ground truth input labels are unavailable, so we use optical flow direction as a proxy. We train the joint IDM across 7 games: Space Invaders, Asterix, Frostbite, Hero, Seaquest, Centipede, and Ms. Pac-Man.

**Joint training clusters by action, not game.** Single-game IDMs learn flow templates that are too specific to generalize. Joint training forces the model to find features that work across all seven visual styles simultaneously. Figure 2 shows the result: embeddings cluster by action class (LEFT, NEUTRAL, RIGHT) with frames from all games distributed across each cluster.

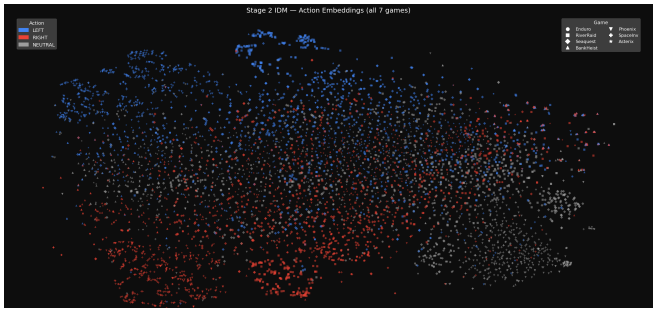


Figure 2: t-SNE of joint IDM embeddings from 7 Atari games, colored by action class: LEFT (blue), RIGHT (red), NEUTRAL (gray). Clusters form by action rather than by game.

**Predicting which game pairs align.** We compute cross-game KNN accuracy: for each game, we embed its validation frames and check whether nearest neighbors from other games share the same action label. This produces the $7 \times 7$ similarity matrix in Figure 3. Games with similar visual dynamics (e.g., Space Invaders ↔ Asterix, both horizontal-scroll games) show high cross-game KNN accuracy, while visually dissimilar pairs (e.g., Ms. Pac-Man ↔ Frostbite) show low accuracy. This metric predicts which game pairs will benefit from joint training *before running the experiment.*

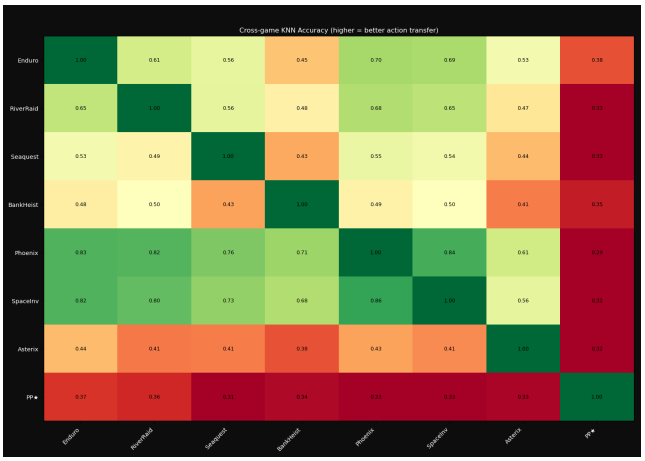


Figure 3: Cross-game KNN accuracy matrix for 7 Atari games. Higher values indicate stronger action alignment, predicting which pairs benefit from joint training.

## 4.3 First-Person Games

**Setup.** First-person games are the hardest domain for cross-game alignment. In racing and Atari games, actions produce localized visual effects (a road edge shifts, a sprite moves). In first-person games, camera rotation produces *global* optical flow across the entire frame, making the spatial templates that worked for racing less effective. We use Minecraft VPT [2] (1.84M frames of human exploration gameplay) and CS:GO Deathmatch (400 sessions of human combat gameplay), both with ground truth keyboard and mouse labels.

**Movement keys align well.** Despite the domain gap between a blocky exploration game and a realistic combat shooter, the joint IDM achieves **85.2%** on Minecraft W-key and **77.8%** on CS:GO W-key (Table 2). This works because forward movement produces a consistent visual signature across both games: objects near the bottom of the frame expand outward, regardless of art style. Figure 4 shows the model correctly predicting W, D, and camera-left simultaneously on a CS:GO frame.

**Camera direction: the calibration gap at scale.** Camera direction is where the calibration gap is most visible. Zero-shot accuracy is 54.5% (Minecraft) and 56.1% (CS:GO) — barely above chance. But a linear probe on the same embeddings reaches 71.9% and 63.9%. The information is there; the decision boundary is wrong.

The root cause is the same as in racing, but ampli-

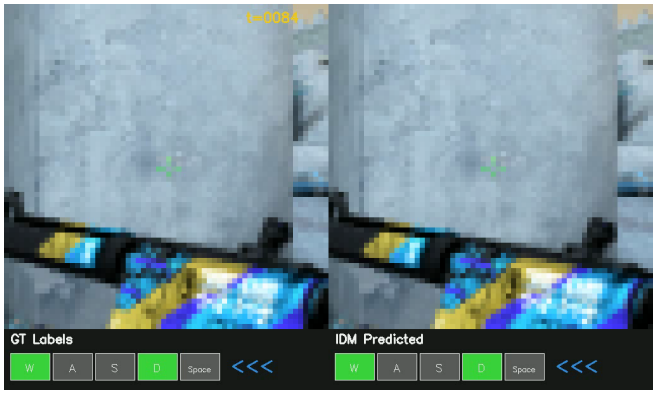


Figure 4: Joint IDM predictions on CS:GO. Left: ground truth. Right: IDM predictions. W, D, and left-camera are all correctly identified.

Table 2: Joint IDM accuracy on first-person games.

| Action | Minecraft | CS:GO |
|---|---|---|
| W (forward) | 85.2% | 77.8% |
| A (left strafe) | 68.4% | 71.3% |
| D (right strafe) | 72.1% | 69.8% |
| Mouse dir. (zero-shot) | 54.5% | 56.1% |
| Mouse dir. (calibrated) | 71.9% | 63.9% |
| Chance | 33.3% | 33.3% |

fied. Minecraft players make smooth, exploratory camera rotations; CS:GO players make fast flicks with long neutral periods. What counts as "neutral" is at a different flow magnitude in each game. With 50 labeled frames per class, the boundary can be repositioned and the full accuracy is recovered.

# 5 The Calibration Gap

Across all three domains, we observe the same failure pattern: the IDM's representation captures the right information, but its decision boundary is biased toward the training distribution of whichever game dominates.

In racing, this shows up as a 15.4 pp gap between the joint IDM (58.3% on Enduro) and a model calibrated with 20% of Enduro's labels (73.7%). In first-person games, it shows up as the gap between zero-shot camera accuracy (54–56%) and calibrated accuracy (64–72%).

**When does it appear?** The calibration gap is most severe when:

1. The action's visual signature is global rather than local (camera rotation fills the entire frame, unlike steering which is concentrated at road edges).
2. The magnitude of the signature varies across games (Minecraft's smooth rotations vs. CS:GO's flick shots).
3. One game dominates the training set, biasing the neutral threshold.

**When is it absent?** The gap is small or zero when the action produces a spatially localized, directionally consistent signature — like forward movement producing bottom-of-frame expansion flow in both Minecraft and CS:GO. In such cases, the visual effect looks nearly identical across games, and no per-game calibration is needed.

**Closing the gap without labels.** The calibration gap can be closed without any labeled data by estimating each game's thresholds from its optical flow distribution. We observe that in most gameplay, roughly two-thirds of frames are "neutral" (no active turning), and the remaining third contains actual turns. By computing flow magnitudes over 1000 unlabeled frames and setting the neutral threshold at the 67th percentile, we recover **91.1%** accuracy on Pole Position — within 0.3 pp of the supervised ceiling. This suggests that the calibration gap, while real, is a solvable engineering problem rather than a fundamental limitation.

# 6 Future Directions

**Generalizing auto-calibration.** The flow-based threshold estimation works well on Pole Position but needs validation across first-person games, where the optimal percentile may differ. Adapting the threshold automatically to each game's flow distribution is an open problem.

**Broader game domains.** Our analysis covers racing, Atari, and first-person games — all domains where actions produce continuous visual motion. Strategy games (StarCraft, Dota 2 [9]) present a different challenge: actions are discrete, sparse, and often invisible in the frame (e.g., selecting a unit). Testing whether the alignment mechanism extends to such domains would clarify its generality.

# 7 Conclusion

We have shown that a jointly-trained IDM produces semantically aligned action embeddings across games with similar visual dynamics. The alignment is confirmed by t-SNE visualizations showing cross-game interleaving within action clusters, validated against ground truth player input labels. When alignment partially fails — as in camera direction for first-person games — the failure is a calibration gap, not a representation gap: the same embeddings that fail zero-shot succeed with minimal calibration. This gap can be closed without labels using optical flow statistics alone, suggesting that true zero-shot cross-game action transfer is within reach for world models.

# References


[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. DIAMOND: Diffusion for world modeling — visual details matter in atari. In *Advances in Neural Information Processing Systems (Spotlight)*, 2024.

[2] Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795*, 2022.

[3] David Brandfonbrener, Ofir Nachum, and Joan Bruna. Inverse dynamics pretraining learns good representations for multitask imitation. In *Advances in Neural Information Processing Systems*, 2023.

[4] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

[5] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, 2003.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.

[7] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *International Conference on Machine Learning*, 2025.

[8] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[9] OpenAI, Christopher Berner, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[10] Nemanja Rašajski, Chintan Trivedi, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. BehAVE: Behaviour alignment of video game encodings. In *Computer Vision – ECCV 2024 Workshops*, volume 15624 of *Lecture Notes in Computer Science*. Springer, 2024.

[11] Guangxiang Ye, Zichen Lu, Zhilin Huang, Simon Lai, Fuwei Li, and Jian Yao. Learning latent actions to plan with world models. In *International Conference on Learning Representations*, 2023.

[12] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S. Muller, Jake A. Whritner, Luxin Zhang, Mary M. Hayhoe, and Dana H. Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6811–6820, 2020.