Cross-Game Semantic Alignment of Latent Action Representations*

Feolu Kolawole , Khizer Khaderi
Human Perception Lab, Stanford University
flukol@stanford.edu, kkhaderi@stanford.edu

Abstract

We study few-shot and zero-shot cross-domain adaptation of action representations: can a single inverse dynamics model (IDM) trained on multiple visually-disparate games generalize to held-out targets with minimal supervision?

Across three domains (racing, Atari, and first-person games), a jointly-trained IDM clusters embeddings by action class rather than game identity. When transfer partially fails, we show it is a misplaced decision boundary, not a missing feature — what we call the *calibration gap*. The gap is closeable: 50 labeled frames per class are enough to compute action centroids in the joint embedding that recover **91.1%** accuracy on Pole Position, within 0.3pp of a fully-supervised single-game classifier trained on ∼30,000 labels (91.4%).

Extending the framework to held-out games, an RGB IDM transfers only weakly (+4.5pp over chance), but replacing the input with appearance-invariant optical flow more than doubles the transfer (+9.7pp on held-out games; +14.8pp on first-person cross-game evaluation, where RGB is at chance). The open-set weakness of pixel IDMs is an input-representation choice, not an intrinsic limit.

1 Introduction

Unsupervised world models [Bruce *et al.*, 2024; Hafner *et al.*, 2023; Alonso *et al.*, 2024] 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. But do these learned actions have intuitive semantic meaning? If a model watches gameplay from two different racing games, does it learn that "steer left" is the same action in both, or does it memorize two unrelated patterns?

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 is a domain-adaptation question with limited resources: we want one model to work on many visually-disparate environments without per-environment supervision.

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 the supervised accuracy of a single-game classifier via centroid calibration on the joint embedding.

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 — a form of unsupervised domain adaptation.
- 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 few-shot centroid-calibration method: 50 labels per class on the frozen joint embedding recover 91.1% accuracy on Pole Position, within 0.3 pp of a fully-supervised single-game classifier (91.4%).

2 Related Work

Latent action world models. Genie [Bruce *et al.*, 2024] learns latent actions from unlabeled video via a learned action tokenizer. LAPO [Ye *et al.*, 2023] jointly trains an IDM and a forward dynamics model so that latent actions must be

*Accepted at the **IJCAI 2026** 4th International Workshop on Generalizing from Limited Resources in the Open World (GLOW).

both predictable from frame pairs and useful for generation. AdaWorld [Gao *et al.*, 2025] 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. [2023] showed that IDM pretraining outperforms other self-supervised objectives for learning transferable representations. VPT [Baker *et al.*, 2022] 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 [Rašajski *et al.*, 2024] 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 [Ganin *et al.*, 2016]) is the standard approach for cross-domain feature alignment, alongside CORAL [Sun and Saenko, 2016] (covariance matching) and MMD-based methods [Gretton *et al.*, 2012]; we measure DANN as a baseline and find it transfers negatively (28.3% on Enduro for a PP-trained DANN, below 33.3% chance), motivating our joint-training approach.

Few-shot and zero-shot domain adaptation. Few-shot DA spans prototypical methods [Snell *et al.*, 2017], FADA [Motiian *et al.*, 2017], and meta-learning approaches. Our 50-label centroid-calibration result is closest to FADA's regime, but operates entirely on a frozen joint encoder rather than fine-tuning, more in the spirit of confidence-recalibration methods [Guo *et al.*, 2017].

Latent-action analysis. Schmidt et al. [2025] provide a theoretical analysis of what latent action models actually learn, showing that LAM features can be decomposed into controllable and noise components. Our discriminative IDM avoids one of their identified failure modes (action-conditioning collapse under reconstruction objectives) by training on supervised action labels rather than predicting future frames. Our cross-game KNN matrix (Fig. 3) is closely related to their LLO transferability metric.

Game data sources. We use Atari-HEAD [Zhang *et al.*, 2020] (17-game human Atari demonstrations), the VPT contractor dataset [Baker *et al.*, 2022] (Minecraft human gameplay, 1.84M frames at 20fps), and the Counter-Strike: GO Deathmatch dataset [Pearce and Zhu, 2022] (400 sessions, 16.7fps). Pole Position and Enduro labelled data were collected by the authors via real-time keypress logging during human play.

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 [Farnebäck, 2003] (per-pixel apparent motion 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.

Calibration gap (definition). We formalize the central concept of this paper: for a target domain T and a joint IDM trained on source domains S, the **calibration gap** is

$$\mathrm{CG}(T) \;=\; \mathrm{Acc}_{\text{probe}}(T) \;-\; \mathrm{Acc}_{\text{joint}}(T),$$

where $\mathrm{Acc}_{\text{probe}}(T)$ is the balanced accuracy of a linear probe trained on the joint IDM's penultimate features for T, and $\mathrm{Acc}_{\text{joint}}(T)$ is the joint model's zero-shot accuracy on T. A non-zero gap means the representation contains the information needed to classify T, but the joint model's decision boundary is mis-placed for T. We measure CG in every domain.

Implementation details. Frame inputs are $64\times64\times3$. The IDM CNN has channels $\{64, 128, 256, 256\}$ with 3×3 kernels and stride-2 down-sampling, producing a 16×16 spatial latent grid. A 1×1 convolutional head maps each spatial cell to per-class logits, which are global-average-pooled to a single distribution over action classes. We train with cross-entropy plus a small cross-game supervised-contrastive auxiliary loss (SupCon [Khosla *et al.*, 2020], weight λ=0.1) that pulls same-action embeddings together across games, Adam (lr $= 10^{-3}$), batch size 64, and 8000–10000 steps. Frame stride between paired frames is set per domain: 7 for racing and Atari, 2 for first-person games (selected by validation accuracy). Hflip augmentation (p=0.5) is applied for racing and Atari with corresponding label swap. All experiments use a single A100 / L40S GPU; training takes 1–3 hours per model.

4 Experiments

We apply the framework to three domains of increasing difficulty: racing, Atari, and first-person games.

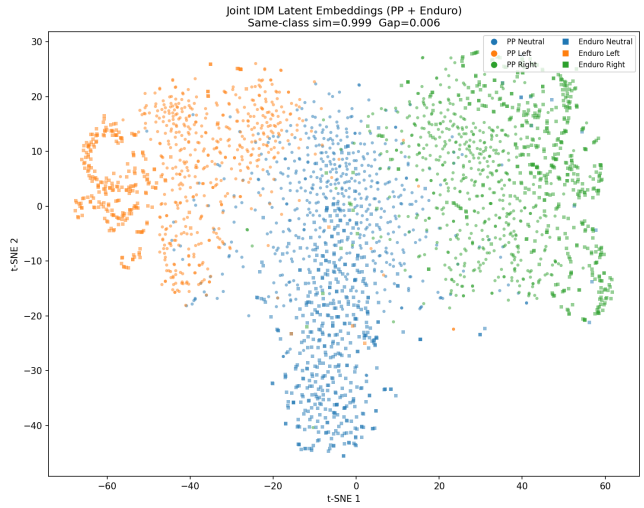

Figure 1: t-SNE of joint IDM embeddings for Pole Position (circles) and Enduro (triangles), colored by steering direction. Frames from both games are interleaved within each action cluster, confirming semantic alignment.

Table 1: Balanced accuracy on racing games. **Calibrated (20% labels)**: joint IDM with the classifier head re-fit on 20% of Enduro labels (frozen encoder).

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)	76.6%	59.3%
Joint IDM v3 (+hflip)	**80.2%**	**58.3%**
Calibrated (20% labels)	—	73.7%

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.

The in-distribution calibration gap (first appearance). Because the joint IDM is trained on both PP and Enduro labels, evaluation on either is in-distribution. It reaches 58.3% on Enduro, but fine-tuning with just 20% of Enduro's labels (a probe-only adaptation on the same encoder) pushes this to 73.7% — a 15.4 pp in-distribution calibration gap by the definition above. Crucially, the t-SNE clusters are already well-separated; the gap is not in the representation but in the decision boundary. Enduro has $10\times$ 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 in-distribution phenomenon reappears in first-person games at larger scale (Section 4.3), and the open-set version is treated separately in Section 4.2.

The gap is calibration, not capacity. A 20%-label probe on Enduro lifts the joint IDM from 58.3% to **73.7%** — a 15.4pp gap that closes when the decision boundary is allowed to adapt, with no change to the encoder. The same recalibration pattern recurs in first-person games at larger scale (Section 4.3).

4.2 Atari Games

Setup. We use the Atari-HEAD dataset [Zhang *et al.*, 2020], 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, Phoenix, Enduro, River Raid, Seaquest, Bank Heist, and Asterix.

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. Per-game balanced accuracy (Table 2) ranges from 50.0% to 89.7% across the 7 games, with a mean of 71.3%. Games with strong horizontal-flow signatures (Space Invaders, Phoenix) reach 88%+, while games with heavy vertical action (Asterix, Bank Heist) suffer from 3-class label collapse: their UP/DOWN movements get folded into NEUTRAL, polluting the neutral cluster.

Table 2: Per-game balanced accuracy of the joint Atari IDM (single weights, 7 games; mean 71.3%). **UP/DOWN freq.**: fraction of player frames pressing UP or DOWN in each game.

Game	Bal Acc	LEFT/RIGHT	UP/DOWN freq.
Space Invaders	0.897	0.901 / 0.911	∼0%
Phoenix	0.877	0.888 / 0.888	0.2%
Enduro	0.773	0.818 / 0.831	2.3%
River Raid	0.764	0.731 / 0.710	12.6%
Seaquest	0.612	0.645 / 0.634	23.9%
Bank Heist	0.565	0.670 / 0.517	42.9%
Asterix	0.500	0.631 / 0.535	25.1%
Mean	**0.713**	—	—

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 7×7 similarity matrix in Figure 3. Games with similar visual dynamics (e.g., Space Invaders $\leftrightarrow$ Asterix, both horizontal-scroll games) show high cross-game KNN accuracy, while vi-

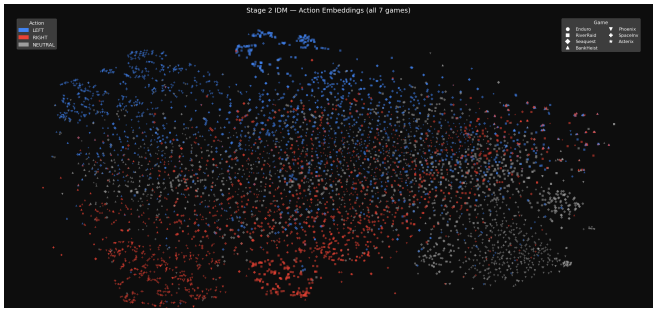

Figure 2: t-SNE of joint IDM embeddings from 7 Atari games, colored by action class: LEFT (blue), RIGHT (red), NEUTRAL (gray). Clusters form by action rather than by game.

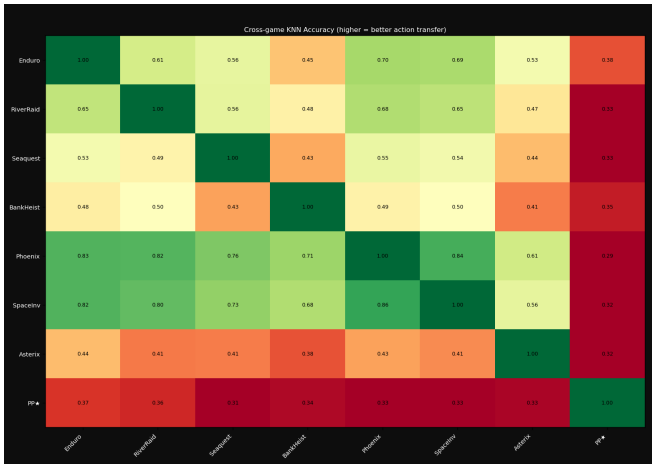

Figure 3: Cross-game KNN accuracy matrix for 7 Atari games. Higher values indicate stronger action alignment, predicting which pairs benefit from joint training.

sually 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*.

Open-set generalization to held-out games. The results so far measure alignment among the 7 games in the joint training pool. We now ask a stronger question: how well does the joint IDM transfer to games it has *never seen* during training? We test this with a leave-one-out (LOO) protocol: for each of the 7 paper games, we retrain the joint IDM on the remaining 6 and evaluate on the held-out target.

Calibration alone is insufficient open-set. Table 3 shows the 9-class story. Held-out accuracy drops to 25.3% (vs. 40.5% in-distribution) — a 15.2pp open-set gap. Adding a frozen-encoder logistic probe trained on N held-out labels per class does *not* close it: on 5 of 7 games even N=200 underperforms zero-shot LOO, and the mean N=50 probe (21.3%) is actually worse than zero-shot (25.3%). When the encoder has not seen a game's visual style, target labels cannot fabricate features the representation never formed. This is a *representation gap*, not a calibration gap — so closing it requires changing the input representation, not the decision boundary.

Table 3: Open-set 9-class probing on held-out Atari games. k: active 9-class actions (chance $= 1/k$). **In-dist**: 9-class accuracy when the game is in the joint pool. **LOO**: zero-shot when held out. N=50, N=200: frozen-encoder logistic probe trained on N held-out labels/class.

Game	k	**In-dist**	**LOO**	N=50	N=200
Space Invaders	3	45.7%	34.6%	34.8%	40.6%
Phoenix	4	50.3%	39.7%	38.0%	45.8%
Enduro	7	56.8%	30.4%	14.8%	20.5%
River Raid	9	25.9%	16.7%	15.8%	15.2%
Seaquest	9	40.7%	19.7%	16.0%	16.4%
Bank Heist	9	31.5%	17.2%	16.1%	16.8%
Asterix	9	32.8%	18.8%	13.6%	14.5%
Mean		**40.5%**	**25.3%**	**21.3%**	**24.3%**

Flow input closes most of the representation gap. Motivated by the negative result, we replace the RGB input with dense optical flow ($[\text{flow}_x, \text{flow}_y, \text{flow}_{\text{mag}}]$, Farnebäck) using the same training recipe; no other change. Over 3 seeds, the same cross-game probe roughly *doubles* its gain over chance (+4.5pp → **+9.7pp**; Table 4, RGB vs. Flow columns), with the largest gains on visually clean horizontal-motion games. Adding 50 held-out LRN labels on top of the frozen flow encoder then closes most of the residual gap (Flow+50 column), except for maze-like games whose dynamics carry no directional flow signature and stay near chance even at N=100. This is not under-optimization: we also tried flow amplification, sign-preserving squaring, global-flow subtraction, direction-only encoding, a two-stream RGB+flow encoder, and a 5–6× larger encoder — none meaningfully improves over base flow (all within ∼2pp). The lever is appearance-invariance, not added capacity.

Table 4: Open-set LRN balanced accuracy on held-out Atari games (chance = 33.3%; Flow columns mean±std over 3 seeds). **LOO 9-cl**: zero-shot 9-class accuracy (Table 3). **RGB xfer**, **Flow xfer**: best probe trained on the OTHER 6 games' LRN labels (no target labels), with RGB and optical-flow inputs respectively. **Flow+50**: flow encoder frozen, logistic head fit on 50 held-out labels/class.

Held-out game	**LOO 9-cl**	**RGB xfer**	**Flow xfer**	**Flow+50**
Space Invaders	34.6%	38.9%	49.5 ±1.0%	56.0 ±1.0%
Phoenix	39.7%	44.1%	**62.2** ±2.0%	61.9 ±2.0%
Enduro	30.4%	36.4%	37.8 ±1.0%	**54.0** ±2.0%
River Raid	16.7%	33.9%	39.4 ±3.0%	**50.1** ±3.0%
Seaquest	19.7%	37.9%	41.1 ±3.0%	**48.4** ±1.0%
Bank Heist	17.2%	36.6%	34.5 ±1.0%	36.4 ±1.0%
Asterix	18.8%	37.0%	36.6 ±1.0%	36.9 ±1.0%
Mean	**25.3%**	**37.8%**	**43.0%**	**49.1%**

Interpretation. These results sharpen the calibration-gap thesis and extend it to the open-set regime. (i) On games in the joint pool, the calibration gap is real and closeable with ≤50 source-domain labels (Section 4.1). (ii) On held-out games, an RGB IDM transfers the coarse action axis only weakly (+4.5pp) because it carries an appearance-induced representation gap. (iii) Making the input appearance-invariant (optical flow) roughly doubles zero-label cross-

game transfer (+9.7pp over chance), and the residual gap on motion-ambiguous games is then a calibration gap that 50 target labels substantially reduce. The open-set weakness of pixel IDMs is therefore largely an input-representation choice, not an intrinsic limit — and the calibration-gap framework predicts what remains once that choice is corrected.

Label-derivation concern. Atari LRN labels are themselves derived from Farnebäck flow (Section 3), so one might worry the flow-input gain reflects label leakage rather than appearance-invariance. Two facts argue against this. First, the labels use a *single global dominant-angle threshold per frame pair*, while the flow-input IDM ingests the full per-pixel $[f_x, f_y, |f|]$ field through a CNN that must learn what those channels mean for action — the two share a sensor, not a decision rule. Second and decisively, the *same* flow-over-RGB advantage ($+14.8$pp mean) recurs in first-person games (Section 4.3, Table 5) where labels are raw keyboard/mouse logs with no flow involvement. The lever generalizes to a domain where the leakage hypothesis cannot apply.

4.3 First-Person Games

Setup. Having established on Atari that replacing the RGB input with optical flow roughly doubles open-set cross-game transfer, we ask whether the same lever generalizes to a harder and more application-relevant domain. First-person games are an especially natural fit: camera rotation produces a near-*global* optical-flow signature that should be maximally appearance-invariant across visually disparate titles, and this domain is simultaneously the hardest cross-game setting we consider (global rotation rather than the localized motion of racing/Atari) and the most directly representative of modern game-AI and world-model deployment, where shared action representations across titles carry operational value. We evaluate the joint IDM on Minecraft VPT [Baker *et al.*, 2022] (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's mean key-press accuracy (**77.5%** Minecraft, **70.0%** CS:GO; Table 5) is within 1–2 points of the single-game upper bound (76.0%, 71.2%) — the joint model pays essentially no in-game cost for sharing weights across both titles. Movement keys transfer because they produce a consistent visual signature across both games (Figure 4).

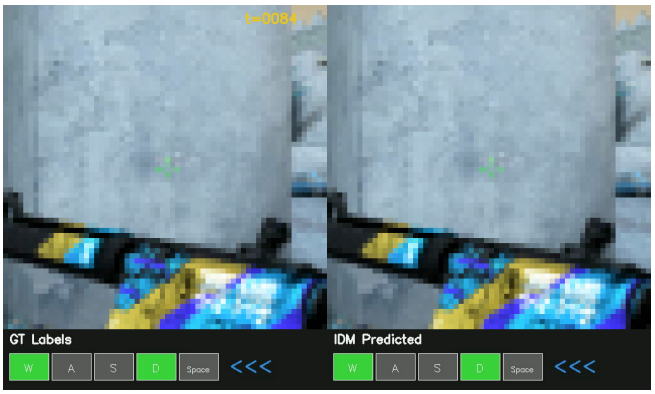

Figure 4: Joint IDM predicts W, D, camera-left on a CS:GO frame.

Table 5: First-person accuracy across three settings. **Single-game**: one IDM per game. **Joint**: one shared IDM trained on both. **Zero-shot cross-game**: trained on one game, evaluated on the other (held-out). Mouse-direction probe = linear probe re-fit on full target labels.

Method	**Minecraft**	**CS:GO**
Chance (mouse dir., 3-class)	33.3%	33.3%
Single-game IDM (per-game weights)		
Key-press mean (5 keys)	76.0%	71.2%
Joint IDM (in-distribution, trained on both games)		
Key-press mean (5 keys)	77.5%	70.0%
Mouse dir.	54.5%	56.1%
Mouse dir. probe (full target labels)	71.9%	63.9%
Zero-shot cross-game (train one, test the other)		
Mouse dir. — RGB IDM	34.2%	33.3%
Mouse dir. — Flow IDM (3-seed)	47.7 $\pm$0.6%	48.4 $\pm$0.4%
Mouse dir. — Flow + 50 target labels	47.4 $\pm$1.3%	42.6 $\pm$0.6%

In-distribution calibration gap (camera direction). The joint IDM is trained on both Minecraft and CS:GO labels, so on each game's val split it is operating *in-distribution*. On mouse-direction, the shared head without target-specific re-calibration reaches 54.5% (Minecraft) and 56.1% (CS:GO); a linear probe re-fit on full target labels on the same encoder reaches 71.9% and 63.9% (Table 5, Joint block). This $\sim$15pp gap is the in-distribution calibration gap at scale — the joint head is biased toward whichever game's mouse-flow distribution dominates training, while the features themselves carry game-specific direction information that the probe recovers.

Zero-shot cross-game transfer: flow recovers most of the gap, but +50 labels do not help further. To test the appearance-invariance lever from Section 4.2 on this harder domain, we train a single-game IDM on one first-person game and evaluate camera direction zero-shot on the other (Table 5, Zero-shot block). An RGB IDM transfers at chance in both directions. Replacing the input with optical flow lifts both directions to roughly 48% ($+14.8$pp mean advantage) — camera rotation produces a near-identical global flow field regardless of game art. Unlike the Atari motion-ambiguous games, adding 50 target labels on the frozen flow encoder does not help and on CS:GO transfer actively

hurts (48.4% → 42.6%, Table 5 last row): the small-sample probe overfits to target noise rather than closing a real residual gap. Flow zero-shot is already at the representational ceiling for the available data, so the residual gap here is a data/representation limit rather than a calibration gap — a distinction the framework makes explicit. **The flow-input advantage thus holds across both zero-shot settings tested: Atari (leave-one-out) and first-person (cross-game)**, complementing the in-distribution few-shot centroid result of Section 4.1.

5 The Calibration Gap

Sections 4.1–4.3 established that an appearance-invariant optical-flow input lets a single IDM transfer actions across visually disparate games: in all three domains — racing, Atari, and first-person — flow input beats RGB. But transfer is seldom perfect, and the residual gap is consistent in form. A model that shares one decision boundary across games inherits a compromise: the games differ in the *magnitude* distribution of their actions (one game's turns or camera motions are larger, faster, or rarer than another's), so the shared threshold for "what counts as a turn" is pulled toward whichever game dominates, leaving the under-represented game systematically miscalibrated even though its features are sound. We call this residual the *calibration gap*. It recurs in both regimes we study — in-distribution (the target is in the joint training pool) and zero-shot (the target is held out) — and in both, it is closed the same way: recover the small amount of game-specific threshold information missing from the shared boundary, from a handful of labeled frames or from unlabeled flow statistics, and re-weight toward the under-valued game.

The in-distribution case is the clearest. In racing, the joint IDM leaves a 15.4 pp gap that target-label calibration recovers; the joint IDM in first-person games shows the same pattern, with recalibration recovering a comparable 8–17 pp. The zero-shot version — where the target game is held out entirely — is treated in the Atari leave-one-out study of Section 4.2.

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 (mouse-movement distributions differ heavily — Minecraft camera motion tends to be smoother, CS:GO more abrupt).
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 in-distribution gap without labels. The in-distribution calibration gap can be closed with very few target labels. Taking the joint IDM's 128-dim pre-head embeddings for Pole Position and computing one centroid per action class from 50 labeled frames per class (150 labels total), nearest-cosine classification recovers **91.1%** balanced accuracy — within 0.3 pp of a fully-supervised single-game classifier trained on the full label set (91.4%). The supervised classifier consumes ∼30,000 labels to reach 91.4%; the centroid needs 150 to reach 91.1%. This ∼200× reduction confirms that the gap is a misplaced decision boundary rather than a missing representation: once the geometry is in place, locating the boundary is cheap.

The embedding space is semantically interpretable. The calibration-gap interpretation assumes the joint IDM's features are *semantically meaningful* — that action classes occupy a structured, interpretable geometry rather than arbitrary clusters. We verify this directly with three measurements on the penultimate features, each a cosine similarity between embedding directions (1.0 = perfectly aligned, 0 = unrelated):

- **Cross-game alignment:** the "steer left vs. right" direction is essentially the same vector in both games — the steering axis learned on Pole Position aligns with the one learned on Enduro at cosine **0.83**.
- **Action factorization:** steering and acceleration lie on near-orthogonal axes (cosine -0.05) — the model encodes them as independent action dimensions rather than entangling the two.
- **Compositionality:** averaging all LEFT-frame embeddings and all RIGHT-frame embeddings and taking their midpoint yields almost exactly the average NEUTRAL embedding (cosine **0.9998**) — "neutral steering" sits halfway between the two extremes, as a semantically meaningful space should.

Together these show the embedding space is not an opaque feature soup but a low-dimensional, interpretable action geometry. This also pins down what the calibration gap is: a translation along the shared steering axis, not a structural mismatch.

6 Discussion

Connection to few-shot and zero-shot domain adaptation. Our results connect to a familiar idea in learning from limited data: *few-shot* adaptation. Once the shared representation is in place, a new game needs only about 50 labeled frames per action to match the accuracy of a fully supervised model trained on ∼200× more data. This succeeds because the calibration gap is a misplaced decision boundary rather than a flawed representation, and locating the boundary just means computing 3 centroids in the existing embedding space — far cheaper than retraining the model end to end.

Why joint IDM training succeeds without adversarial losses. Domain-adversarial methods like DANN [Ganin *et al.*, 2016] explicitly remove domain-discriminative features. We do not use any adversarial loss; the cross-game alignment emerges purely because the supervised classification objective forces the shared CNN to find features that work across

all games. When the same action produces a similar visual signature across games (e.g., bottom-of-frame expansion for forward motion), the model converges on that shared feature naturally. When the signatures differ in magnitude (camera rotation), the model still extracts the right feature but miscalibrates its threshold — the failure mode we identify.

7 Future Directions

Truly label-free calibration. Our 50-label centroid result is few-shot, not zero-shot. A natural next step is recovering the same centroids from purely unlabeled gameplay — e.g., clustering the joint embedding directly, or estimating action priors from optical-flow statistics — so that no target-game annotation is required at all.

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 [OpenAI *et al.*, 2019]) 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.

Beyond games. The cross-game setting is a controlled testbed for cross-embodiment action transfer in robotics. The same calibration-gap phenomenon may arise when transferring action representations between robot platforms with different camera geometries or actuation noise distributions, with the same minimal-supervision fix available.

8 Limitations

Our coverage is uneven across domains: racing and first-person each use only two games, and only Atari approaches a real population (7 games trained, 17 evaluated under the flow-only oracle protocol). The 50-label centroid calibration that recovers Pole Position has been validated only there; whether the same few-shot recipe transfers as cleanly to other in-pool games is an open question. We also focus throughout on discrete 3-class action labels (LEFT/NEUTRAL/RIGHT for steering and camera direction) and leave continuous mouse regression for future work.

The open-set Atari leave-one-out study (Section 4.2) extends the calibration-gap thesis to held-out games once the input is appearance-invariant, but two of the seven held-out games (Bank Heist, Asterix) stay near chance even with 100 target labels: their maze-like dynamics yield no clean directional flow signature — a representation limit, not a calibration gap. All main results are multi-seed (racing and first-person IDMs at $N{=}3$; open-set Atari LOO at $N{=}3$).

9 Conclusion

A jointly-trained IDM produces semantically aligned action embeddings across visually disparate games. When alignment partially fails, the failure is a calibration gap (a decision-boundary shift), not a representation gap (a feature mismatch). The gap can be closed with as few as 50 target labels per class — a single centroid in the joint embedding recovers 91.1% accuracy on Pole Position, within 0.3 points of a fully-supervised single-game classifier (91.4%) trained on $\sim$200$\times$ more data. Combined with the flow-input result on the open-set Atari and first-person settings, this shows that few-shot domain adaptation is broadly tractable for cross-game action representations.

References

[Alonso *et al.*, 2024] 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.

[Baker *et al.*, 2022] 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.

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

[Bruce *et al.*, 2024] 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.

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

[Ganin *et al.*, 2016] 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.

[Gao *et al.*, 2025] 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.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

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

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Motiian *et al.*, 2017] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 2017.

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

[Pearce and Zhu, 2022] Tim Pearce and Jun Zhu. Counter-strike deathmatch with large-scale behavioural cloning. In *IEEE Conference on Games (CoG)*, 2022.

[Rašajski *et al.*, 2024] 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.

[Schmidt and others, 2025] Dominik Schmidt et al. What do latent action models actually learn? *Advances in Neural Information Processing Systems*, 2025.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

[Ye *et al.*, 2023] 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.

[Zhang *et al.*, 2020] 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.