QUERYPRINT: MEASURING BEHAVIORAL SIMILARITY WITH COUNTERFACTUAL FUTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning transforms observations into features for prediction and control. For control, a useful notion of state similarity asks whether the same policies lead to similar futures. Existing behavioral and predictive methods learn state relationships from trajectories, which show one future of each state and rarely the futures of different states under the same policies. We introduce QueryPrint, which restores saved states in resettable environments, executes fixed query policies, and summarizes the resulting counterfactual futures into fingerprints. Distances between fingerprints define a query-relative behavioral target, which we test for action dependence, reliability, and non-degeneracy and use to supervise a compact encoder. Across 55 Atari games, this encoder captures the measured geometry in its native distances and generalizes to unseen states and query policies. In contrast, the 9 learned objectives and 12 frozen pretrained representations we evaluate show little of this geometry in their native distances. Heads trained on the same target recover much of it from frozen features and reach the encoder's level at matched capacity. The measured distance also predicts differences in lives lost under independently trained PPO agents, while a learned world model recovers the geometry only partially. Replayed futures provide a directly measured target for learning and evaluating behavioral geometry, and show that information recoverable from features need not appear in native distances.

1 INTRODUCTION

Representation learning transforms high-dimensional observations into features that support prediction and control. In reinforcement learning, these features can be learned through interaction with an environment or obtained from pretrained visual encoders that are kept frozen (Nair et al., 2022; Ma et al., 2023; Majumdar et al., 2023; Assran et al., 2025). Agents use the resulting state representations to predict outcomes and select actions (Li et al., 2006).

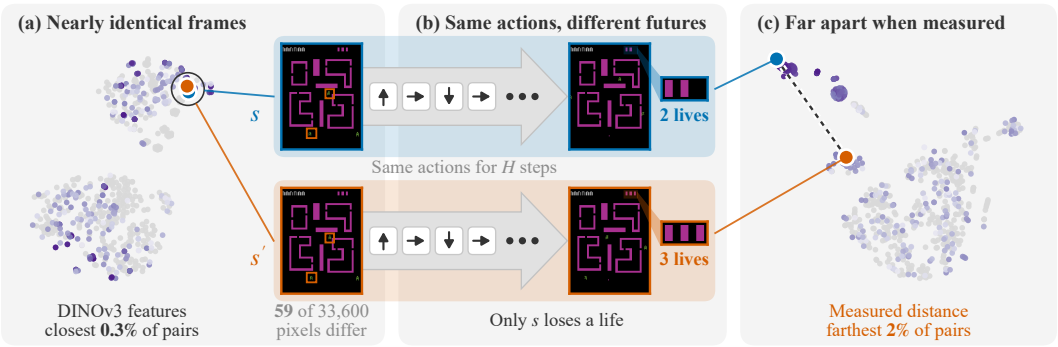


Figure 1: **Similar-looking states can respond differently to the same actions.** The *Venture* states s and s' look nearly identical, and DINOv3 places them closer than almost any other pair of the 600 states. When the same actions are replayed, however, s loses a life and s' does not, and the measured behavioral distance places the pair among the farthest apart. In the t-SNE maps, darker states lose a life sooner under the test queries. Action keys are schematic.

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

To support action selection, a representation should retain information about the consequences of the agent's actions. How two states look, however, does not determine whether they respond similarly to the same actions. In Figure 1, two *Venture* states look alike, yet under the same actions one loses a life and the other does not. This motivates a behavioral notion of state similarity, under which states are similar when the same policies lead to comparable futures rather than merely when they look alike. A representation should therefore place behaviorally similar states close together and dissimilar states apart. Bisimulation metrics formalize this similarity (Ferns et al., 2004; 2006).

Prior work captures structure relevant to behavior with objectives based on rewards and transitions (Zhang et al., 2021; Castro et al., 2021; Shimizu & Tomizuka, 2025), policy behavior (Agarwal et al., 2021a), temporal relationships (Myers et al., 2024; 2025; Park et al., 2026), and latent dynamics (Fujimoto et al., 2025). These methods learn from trajectories, which show one future per state and rarely the futures of different states under the same policies. Downstream evaluations judge a representation by the return of a policy trained on it (Nair et al., 2022; Majumdar et al., 2023). The return also depends on the policy learner, and such evaluations can leave the representation's geometry unexamined (Luo et al., 2025). Whether these representations are behavioral has rarely been tested against measured futures.

In resettable environments such as game emulators, saved states can be restored and continued under the same policies (Bellemare et al., 2013). The futures that define behavioral similarity can then be measured and compared with the distances of any representation. We introduce QueryPrint, which executes a fixed set of simple *query policies* from each saved state and summarizes the resulting changes and outcomes into *fingerprints*. Fingerprint distances define a target distance relative to the query set and independent of the representation being evaluated. We select the target among candidate measurements that pass action dependence, reliability, and non-degeneracy tests on the training queries, and test it again under separate test queries. Figure 2 gives an overview of QueryPrint.

We use the measurement in two ways. It evaluates a representation by comparing the measured geometry with the representation's own distance, which we call the *native distance*, or with the distance of a small *behavioral head* trained on its frozen features. It also supervises a compact QueryPrint encoder, which runs without the emulator once trained. On 55 Atari games, we evaluate 9 learned objectives and 12 frozen pretrained representations, train the encoder, test a learned world model (Zhou et al., 2025), and relate the measured distance to outcomes under independently trained PPO agents. Our contributions are as follows:

- We introduce QueryPrint, which measures a behavioral distance from replayed futures under fixed query policies in resettable environments and tests its validity.
- We evaluate 9 learned objectives and 12 frozen pretrained representations. Their native distances expose little of this geometry, although trained heads recover much of it.
- We supervise representations with the measured distance. A QueryPrint encoder trained from pixels recovers most of the geometry and generalizes to unseen states and test queries, and capacity-matched heads on pretrained features reach the same level.
- We test the measured distance beyond the query set. The distance predicts differences in lives lost under independently trained PPO agents, an ability that only distances supervised by the measurement inherit. A learned world model recovers less than half of its geometry.

2 Related Work

Learning behavioral metrics. A behavioral metric places two states close when they behave alike, whatever they look like. Its classical form is bisimulation (Larsen & Skou, 1991). Two states of a Markov decision process are equivalent when every action gives them the same reward and the same probability of moving into each equivalence class (Givan et al., 2003). Bisimulation metrics relax this equivalence into a distance (Ferns et al., 2004; 2006). Deep methods approximate such distances from sampled transitions. DBC (Zhang et al., 2021) and BS-MPC (Shimizu & Tomizuka, 2025) regress latent distances onto sampled targets, and MICo (Castro et al., 2021) replaces the optimal coupling with an independent one, so that a policy-dependent distance can be estimated from samples. Other objectives redefine what counts as behaving alike. PSE (Agarwal et al., 2021a) compares a reference policy's action distributions along trajectories, and temporal approaches encode goal-reaching relationships through contrastive successor features (Myers et al., 2024), quasimetric

networks (Myers et al., 2025), or dual goal representations (Park et al., 2026). These objectives and the self-supervised baselines we compare need not induce the same geometry, and their trajectory data rarely provide matched continuations from different states under common policies. We instead measure a query-relative distance from replayed futures, test its validity, and use it to supervise and evaluate representations. Predictive state representations (James & Singh, 2004) also use resets and test responses, but to model dynamics rather than to supervise perception.

Pretrained visual representations for control. Many agents reuse frozen pretrained visual encoders and train policies on their features, using either general-purpose image and video models (Oquab et al., 2024; Tschannen et al., 2025; Assran et al., 2025; Siméoni et al., 2026) or encoders pretrained for control with time-contrastive, value-implicit, or masked objectives (Nair et al., 2022; Ma et al., 2023; Majumdar et al., 2023). Such encoders are often judged by policy return, which also depends on the policy learner and task. Probing benchmarks instead test whether annotated state variables can be decoded from frozen features (Anand et al., 2019), but do not directly assess whether feature distances reflect behavioral similarity. Our measurement complements these evaluations by testing whether the native distance of frozen features reflects the measured relationships and whether a behavioral head can recover them from the same features.

World models and their evaluation. World models learn environment dynamics to support planning or policy learning (Hafner et al., 2025; Valevski et al., 2025; Zhou et al., 2025) and are evaluated through prediction quality, the return of policies trained with them, or closed-loop task success (Zhang et al., 2026). In domains formalized as deterministic finite automata, Vafa et al. (2024) show that generative models can pass conventional diagnostics while failing tests of the underlying world structure. We provide a complementary behavioral test by comparing distances between fingerprints of imagined rollouts with those measured in the emulator under the same queries.

3 PRELIMINARIES

We consider a Markov decision process with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel P, and reward r. An observation function maps a state s to an observation o of one or more recent frames, and a representation f compares two states by the distance between the embeddings of their observations. In games a small visual detail can decide what happens next, so states with similar frames can have different futures, as in Figure 1.

Resettable environments. Emulators and physics simulators can save a state and restore it later, so the same state can be continued under any policy π for any horizon H. For a set of query policies $\Pi = \{\pi_1, \ldots, \pi_K\}$ and a state s, one can observe repeated draws of the future $\tau_k \sim P^{\pi_k}(\cdot \mid s)$ under every query policy. Replay involves no learning and no policy other than the queries, and the distance defined on the replayed futures can be tested for validity directly. Throughout the paper, behavioral distance is the difference between the counterfactual futures of two states under a fixed set of query policies, measured rather than learned.

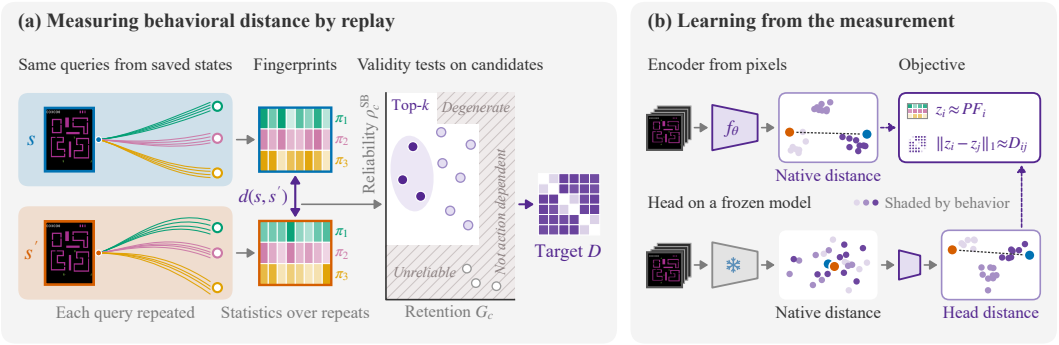


Figure 2: **Overview of QueryPrint.** Replayed futures give a behavioral target for representations. Plots are schematic.

4 QUERYPRINT

As Figure 2 outlines, QueryPrint turns replayed futures into fingerprints and a distance (§ 4.1), tests whether a measurement is valid (§ 4.2), selects the target distance among the measurements that pass (§ 4.3), and uses it to supervise an encoder (§ 4.4).

4.1 FINGERPRINTS

Fix a horizon H, a number of repeats R, and a set of query policies Π, partitioned into training queries Π_{train} and test queries Π_{test}. The query policies are fixed procedural policies, and the measurement requires only that their futures differ. Restoring state s and running query policy π for H steps yields, at every step, a signal of how the world changed, along with the outcome signals that the emulator provides, such as reward and termination. We use three families of change signals, a fixed random projection of the change in emulator memory, in-game events, and a fixed random projection of the change in the frame. A *measurement* $c = (\text{family}, H_c, \gamma_c)$ fixes one family, a horizon $H_c \le H$, and a discount, and $\mathcal{C}$ denotes the grid of candidate measurements. Each measurement summarizes one rollout by the discounted accumulation of the change signal x_t^c of its family, followed by summaries y^c of the outcome signals over the same horizon,

$$u_\pi^c(s) \;=\; \Big[\, \sum_{t=1}^{H_c} \gamma_c^{\,t-1}\, x_t^c \,,\; y^c \Big]. \tag{1}$$

The environment and the queries are stochastic, so each query runs R times, giving R draws of $u_\pi^c(s)$. Let $\phi\big(u_\pi^c(s)\big)$ collect the mean, standard deviation, and 10th and 90th percentiles of every coordinate over these draws. The *fingerprint* of s under a query set Π' concatenates them over queries, and the measured distance is the mean absolute difference between standardized fingerprints,

$$F_{\Pi'}^c(s) = \big[\phi(u_\pi^c(s))\big]_{\pi\in\Pi'}, \qquad d_{\Pi'}^c(s,s') = \frac{1}{\dim F_{\Pi'}^c} \,\big\|\, \bar{F}_{\Pi'}^c(s) - \bar{F}_{\Pi'}^c(s') \,\big\|_1. \tag{2}$$

Here $\bar{F}$ standardizes every coordinate of ϕ by its mean and standard deviation over the states and the training queries, and the same scale is applied under the test queries. Over a set of N states the distances form a matrix $D_{\Pi'}^c \in \mathbb{R}^{N\times N}$. We write F_{train}^c and D_{train}^c for the training queries and F_{test}^c and D_{test}^c for the test queries. Selection and learning use the former, and evaluation uses the latter. A different query set defines a different behavioral distance on the same states, as a different reward defines a different bisimulation metric.

4.2 VALIDITY TESTS

We compute two statistics of every measurement from the training-query fingerprints. Both are properties of the measurement alone, and three tests derived from them decide whether it can serve as a target. The block $\bar{F}_\pi^c(s)$ that query π contributes to the standardized fingerprint splits into a shared part, which is the average future of the state under the query set, and a query-specific part,

$$\bar{F}_\pi^c(s) = M^c(s) + \mathring{F}_\pi^c(s), \qquad M^c(s) = \frac{1}{|\Pi_{\text{train}}|} \sum_{\pi'\in\Pi_{\text{train}}} \bar{F}_{\pi'}^c(s). \tag{3}$$

Both parts are counterfactual quantities, and the distance of Equation 2 compares both. Permuting the query labels within a state reorders the query-specific parts and leaves the shared part unchanged, so only the former can show whether the futures depend on which query acted.

Action dependence. If the futures do not depend on what the query policies do, the distance is not behavioral. Let $\mathring{D}_{\text{train}}^c$ be the distance matrix of Equation 2 computed on the query-specific parts $\mathring{F}_\pi^c$ in place of $\bar{F}_\pi^c$. Let $\tilde{D}_{\text{train}}^{c,\sigma}$ be the same matrix after a permutation σ that reassigns, independently within every state, which query is credited with which recorded future. The *shuffle retention*

$$G_c = \mathrm{Q}_{0.95}\Big\{ \rho_{\mathrm{S}}\big(\tilde{D}_{\text{train}}^{c,\sigma},\, \mathring{D}_{\text{train}}^c\big) \Big\}_\sigma \tag{4}$$

is the 95th percentile, over independent permutations σ, of the Spearman correlation ρ_{S} between the upper triangles of the two matrices. The retention is close to 1 when the futures do not depend on the queries, and it must fall below a threshold τ_{shuffle}.

Reliability. Independent repeats of the measurement must agree. Splitting the R repeats into two halves gives two estimates of D^c_{train}, and their Spearman correlation ρ_c, averaged over splits, estimates the agreement of two half-size measurements. The Spearman-Brown correction $\rho^{\text{SB}}_c = 2\rho_c/(1+\rho_c)$ estimates the agreement of two independent full measurements. This split-half *reliability* is the score that an independent repetition would obtain. Reliability below a floor $\tau_{\min}$ disqualifies a measurement, and above the floor it is the reference to which every score is compared.

Non-degeneracy. A distance can be perfectly reproducible for reasons unrelated to behavior, for example when the futures are dominated by a timer or by scripted motion. A measurement is rejected when its reliability exceeds τ_{rel} while its retention is at or above τ_{weak}.

The set of valid measurements is therefore

$$\mathcal{V} = \big\{\, c \in \mathcal{C} \;:\; G_c < \tau_{\text{shuffle}} \;\wedge\; \rho^{\text{SB}}_c \geq \tau_{\min} \;\wedge\; \neg\big(\rho^{\text{SB}}_c > \tau_{\text{rel}} \wedge G_c \geq \tau_{\text{weak}}\big) \,\big\}. \tag{5}$$

4.3 Target distance

Every candidate in the grid $\mathcal{C}$ of signal families, horizons, and discounts is tested as in § 4.2, and the valid candidates are ranked by their action dependence $1 - G_c$. The target fingerprint concatenates the training-query fingerprints of the k highest-ranked valid candidates, shared and query-specific parts alike, and the target distance is the distance of Equation 2 on it,

$$F_{\text{train}}(s) = \big[F^c_{\text{train}}(s)\big]_{c \in \text{top}_k(\mathcal{V})}, \qquad D_{ij} = \frac{1}{\dim F_{\text{train}}} \big\| \bar{F}_{\text{train}}(s_i) - \bar{F}_{\text{train}}(s_j) \big\|_1. \tag{6}$$

Under the test queries, the same k measurements give F_{test} and D_{test}. Combining the k measurements raises reliability. The procedure declares an environment unmeasurable when fewer than k candidates are valid, rather than yielding an invalid target. A representation is evaluated by the Spearman correlation between its own pairwise distances on the same states and D_{test}. The query policies and the full procedure are given in Appendices A and B.

4.4 Learning from measured distances

An environment provides N training states with observations o_i, standardized target fingerprints $F_i = \bar{F}_{\text{train}}(s_i)$, and a target distance matrix D. We train the *QueryPrint encoder* f_θ, which maps an observation to an embedding $z = f_\theta(o) \in \mathbb{R}^m$. The embedding plays two roles at once. It predicts a fixed random projection $PF_i \in \mathbb{R}^m$ of the fingerprint. Its pairwise mean absolute distances, as in Equation 2, are aligned with the measured distances over every pair of a minibatch B,

$$\mathcal{L}(\theta) = \frac{1}{|B|} \sum_{i \in B} \bar{\ell}\big(z_i - PF_i\big) + \frac{1}{\binom{|B|}{2}} \sum_{i<j \in B} \ell\Big(\tfrac{1}{m}\|z_i - z_j\|_1 - D_{ij}\Big), \tag{7}$$

where ℓ is the Huber loss with unit threshold and $\bar{\ell}$ its mean over coordinates. The alignment term supervises the geometry directly, and the regression term anchors the embedding to the fingerprint coordinates, which keeps it stable where alignment alone is not. The number of updates $T^\star$ is selected on a validation subset of the training states, and the encoder is then reinitialized and trained on all N states for $T^\star$ updates. At inference the encoder runs without the emulator. The overall cost is dominated by the $N\,|\Pi|\,R\,H$ emulator steps per environment.

5 Experiments

The experiments on 55 Atari games address three questions. Do current representations, learned from single trajectories or pretrained, expose the measured geometry, and what recovers it from their features? Does measured supervision carry over to unseen queries, states, and policies? How far does a learned world model reproduce the counterfactual futures of the emulator? Implementation details are provided in Appendices A to E, and Appendix F reports the geometry by game split.

5.1 Setup

Benchmark and measurement. All experiments use the Arcade Learning Environment (Bellemare et al., 2013) with sticky actions and the full action set (Machado et al., 2018). We use 55 of its

57 games, since *Road Runner* yields too few states at some stages of episode progress and *Gravitar* fails the validity tests. The games are split into 37 training games, used to develop the method, and 18 test games. Unless stated otherwise, every number aggregates all 55 games. For each game we sample 600 states uniformly over five stages of episode progress and store them as emulator snapshots with their four most recent frames. Each state is measured for 600 steps with 4 repeats under six procedural query policies, three for training and three for testing.

Representations. We compare two groups on the same states. *Learned objectives* are trained from scratch, one model per game, with the same IMPALA encoder (Espeholt et al., 2018), the same training states, and the same validation rule and update budget as the QueryPrint encoder. They are BS-MPC (Shimizu & Tomizuka, 2025), CMD (Myers et al., 2024), CURL (Laskin et al., 2020), DBC (Zhang et al., 2021), MICo (Castro et al., 2021), MR.Q (Fujimoto et al., 2025), PSE (Agarwal et al., 2021a), ST-DIM (Anand et al., 2019), and TMD (Myers et al., 2025), following the implementations published by their authors. The objectives learn from the stored transitions of the 600 states, and the QueryPrint encoder learns from their measurement, which costs 8.64M emulator steps per game, about 4% of a standard 200M-frame Atari run (Machado et al., 2018). The comparison holds the states, the network, and the updates fixed and varies the source of supervision. *Pretrained representations* are evaluated frozen with the preprocessing prescribed by their authors. They are DINOv2 (Oquab et al., 2024), DINOv3 (Siméoni et al., 2026), SigLIP 2 (Tschannen et al., 2025), V-JEPA 2 and V-JEPA 2.1 (Assran et al., 2025), VLM2Vec-V2 (Meng et al., 2026), and the control-pretrained R3M (Nair et al., 2022), VC-1 (Majumdar et al., 2023), and VIP (Ma et al., 2023). The QueryPrint encoder (1.3M parameters) is the same IMPALA network with a two-layer head.

Scoring. A representation is scored twice on a game with the Spearman correlation of § 4.3 against the test queries, which are unseen during training. The first uses its native distance, the method's own metric or a fixed distance on its features. The second uses a *behavioral head*, a small network on a learned embedding or on a fixed 128-dimensional random projection of frozen features. It is fit on the training states with the objective of Equation 7 under the validation rule of the encoder. Our encoder has no separate head and uses its native distance for both scores. We keep the head small to ask what each representation makes readily available, and § 5.4 tests a wider one. Two further scores go beyond the geometry of the training states. One is the same score on 100 additional states of each test game. The other is fingerprint prediction, the Spearman correlation between the test-query fingerprint coordinates and their ridge-regression prediction from the representation, on episodes not used to fit the regression. Four reference values accompany the scores, a randomly initialized encoder, the agreement between the two query sets, the split-half reliability, and fingerprints averaged over identical observations. Every configuration uses three seeds per game. We report the interquartile mean over games and seeds with 95% stratified bootstrap confidence intervals (Agarwal et al., 2021b), and differences from the encoder are paired by game.

5.2 Validity of the measurement

Figure 3 plots the two statistics of § 4.2 for the selected target of every game on the test queries. A valid target exists for all 55 games. On the test queries, which took no part in the selection, every target passes all three tests again, except that the reliability of *Freeway* falls just below the floor. Permuting which query produced which future destroys most of the structure of the query-specific part, and the median retention is below 0.5. Independent halves of the repeats agree above 0.9 in most games. The highest-ranked measurement differs across games in signal family and horizon. A distance built only from the outcome statistics of the fingerprint reaches 0.515 $[0.464,\ 0.562]$ against the target. The target contains the outcome structure and also captures how the queries change the state. See Appendix N for the selected measurements.

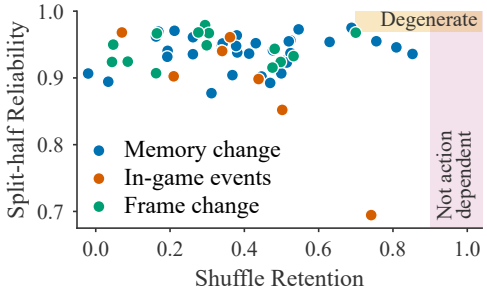


Figure 3: **Validity of the measurement.** Each point is one game on the test queries, and shading marks the regions that the tests reject.

Table 1: **Behavioral geometry of every representation.** Geometry scores use test queries unseen in training, and parameters include the head. Bold marks the best and underline the second.

Representation	Parameters	Input Frames	Behavioral geometry, 55 games[†]		Additional states, 18 games	Fingerprint prediction
			native distance	behavioral head	behavioral head	test queries
Reference						
Random encoder	1.3M	4	0.264 [0.19, 0.35]	0.207 [0.14, 0.30]	0.160 [0.08, 0.39]	0.181 [0.14, 0.23]
Learned objectives (same states, encoder, and update budget)						
BS-MPC	1.3M	4	0.263 [0.19, 0.36]	0.230 [0.16, 0.32]	0.222 [0.11, 0.43]	0.190 [0.15, 0.24]
CMD	1.4M	4	0.200 [0.14, 0.27]	0.511 [0.46, 0.57]	0.406 [0.23, 0.58]	0.190 [0.14, 0.24]
CURL	1.3M	4	0.165 [0.12, 0.23]	0.735 [0.69, 0.77]	0.666 [0.50, 0.75]	0.226 [0.18, 0.28]
DBC	1.3M	4	0.211 [0.15, 0.30]	0.326 [0.24, 0.41]	0.311 [0.13, 0.53]	0.200 [0.16, 0.25]
MICo	1.3M	4	0.197 [0.14, 0.27]	0.208 [0.15, 0.30]	0.243 [0.10, 0.45]	0.226 [0.19, 0.28]
MR.Q	1.3M	4	0.248 [0.18, 0.34]	0.329 [0.25, 0.41]	0.331 [0.16, 0.56]	0.195 [0.15, 0.25]
PSE	1.4M	4	0.130 [0.09, 0.20]	0.566 [0.51, 0.61]	0.530 [0.35, 0.65]	0.200 [0.15, 0.25]
ST-DIM	1.3M	4	0.152 [0.11, 0.21]	0.632 [0.58, 0.68]	0.580 [0.45, 0.68]	0.181 [0.14, 0.23]
TMD	1.5M	4	0.185 [0.12, 0.25]	0.245 [0.19, 0.32]	0.165 [0.06, 0.33]	0.169 [0.13, 0.22]
Frozen visual representations						
DINOv2-B/14	87M	1	0.263 [0.20, 0.34]	0.728 [0.68, 0.77]	0.602 [0.41, 0.73]	0.251 [0.21, 0.30]
DINOv2-L/14	305M	1	0.263 [0.20, 0.34]	0.723 [0.68, 0.77]	0.641 [0.46, 0.74]	0.251 [0.21, 0.30]
DINOv3-L/16	303M	1	0.274 [0.21, 0.36]	<u>0.780</u> [0.75, 0.81]	0.565 [0.45, 0.69]	<u>0.303</u> [0.25, 0.36]
SigLIP 2 So400m/16	428M	1	0.262 [0.20, 0.34]	0.769 [0.73, 0.80]	0.688 [0.57, 0.77]	0.282 [0.23, 0.34]
V-JEPA 2-L (frame)	326M	1	0.253 [0.19, 0.33]	0.707 [0.66, 0.76]	0.599 [0.41, 0.70]	0.252 [0.21, 0.30]
V-JEPA 2-L (clip)	326M	16	0.247 [0.18, 0.33]	0.735 [0.69, 0.78]	0.603 [0.45, 0.73]	0.247 [0.20, 0.30]
V-JEPA 2-g (clip)	1.0B	16	0.235 [0.17, 0.32]	0.741 [0.70, 0.78]	0.654 [0.50, 0.75]	0.251 [0.20, 0.31]
V-JEPA 2.1-L (clip)	305M	16	0.234 [0.18, 0.31]	0.771 [0.73, 0.80]	<u>0.721</u> [0.63, 0.79]	0.275 [0.22, 0.34]
VLM2Vec-V2	2.2B	16	0.261 [0.20, 0.34]	0.771 [0.73, 0.81]	0.653 [0.56, 0.73]	0.258 [0.21, 0.32]
Frozen control-pretrained representations						
R3M ResNet-50	24M	1	0.276 [0.21, 0.36]	0.712 [0.67, 0.76]	0.548 [0.40, 0.66]	0.259 [0.21, 0.31]
VC-1 ViT-L/16	303M	1	0.266 [0.20, 0.35]	0.730 [0.69, 0.77]	0.643 [0.45, 0.75]	0.275 [0.23, 0.33]
VIP ResNet-50	26M	1	<u>0.292</u> [0.23, 0.37]	0.719 [0.68, 0.76]	0.536 [0.39, 0.66]	0.258 [0.21, 0.31]
QueryPrint encoder	1.3M	4	**0.839** [0.82, 0.86]	**0.839** [0.82, 0.86]	**0.743** [0.65, 0.82]	**0.564** [0.50, 0.60]

[†]References, not ranked. Query agreement 0.896 [0.88, 0.91], split-half reliability 0.941 [0.93, 0.95], observation aliasing 0.964 [0.95, 0.97], best common distance 0.294, emulator memory on the 18 test games 0.488 [0.37, 0.61].

5.3 Main results

No baseline representation exposes the measured geometry in its own distance. In Table 1, the native distances of the 12 frozen models and the 9 learned objectives remain at or near the level of a randomly initialized encoder. The QueryPrint encoder reaches 0.839 with its native distance, 94% of the agreement between the two query sets. Neither four input frames nor standardized features raise the frozen models' native scores, and the highest native score under any common distance is 0.294. Aliasing does not explain these scores, since fingerprints averaged over identical observations still reach 0.964. The emulator state does not expose the geometry either. On the 18 test games, the distance between the memory of two states reaches 0.488, below the head of every frozen model.

Frozen features hold the geometry, and a small trained head recovers most of it. The behavioral head raises every frozen model above 0.7, and DINOv3 scores highest, at 0.780. Pretraining thus brings a head of 197,632 parameters close to the encoder trained on measured futures. The DINOv3 head still scores below the encoder on 44 of 55 games, and a wider head narrows the gap (§ 5.4). The small head matches the encoder where the geometry is a simple function of the appearance of the frame, as in *Bank Heist*, *Breakout*, *Hero*, and *Riverraid*, in which even the head on a random encoder scores high. Its largest gaps are in games whose futures depend on small objects that a pooled feature of one frame resolves coarsely, such as the shots in *Assault* and *Battle Zone*, the cars in *Enduro*, or the ball in *Pong*.

Objectives that learn from one future per state recover only part of the geometry. Through the head, the contrastive objectives CURL, ST-DIM, PSE, and CMD recover more of it than the others. Their native scores are among the lowest in the table, below the random encoder, since each arranges its embedding by its own target quantity and not by the futures. The bootstrapped objectives MR.Q, DBC, MICo, and BS-MPC remain near the random encoder. Their embeddings collapse to a few effective dimensions, a failure reported for DBC under sparse rewards (Kemertas & Aumentado-Armstrong, 2021). CURL, the strongest at 0.735, scores below the encoder on 50 of 55 games. No objective comes close to the encoder through its head even with fixed budgets of up to 8,000 updates

on 20 games, and on five games, training on four times as many states leaves the score of CURL unchanged.

Beyond the training states, only distances supervised by the measurement expose the geometry. On the 100 additional states of each test game, the native distance of our encoder keeps most of its score, at 0.743, while every other native distance remains at the random-encoder level. Every head scores lower on these states than on the training states of the same games. DINOv3 drops the most, and V-JEPA 2.1 comes closest to the encoder, on par with it. The encoder also scores highest in fingerprint prediction, which tests how well a representation predicts the futures themselves under the test queries. Appendices D, H, I, and N report further results.

5.4 Analysis

The QueryPrint encoder recovers the part of the test-query geometry that the training queries determine. In Figure 4, its score rises and falls with the agreement between the training and test queries of a game. It is within 0.05 of this agreement in most games. *Freeway* is the one game whose query sets agree below 0.75, because no training query holds up or down, the only actions that move its player. There the score equals the agreement. The agreement acts as an empirical ceiling, which the encoder exceeds in 4 of the 55 games. More varied training queries may raise it.

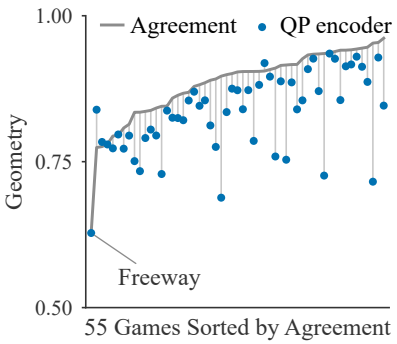
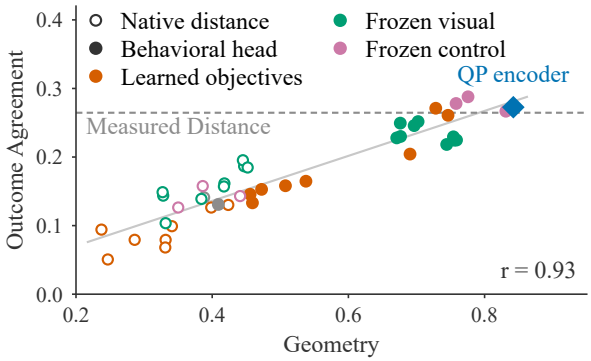


Figure 4: **Geometry against query agreement.** The QueryPrint (QP) encoder on every game, sorted by the agreement of its two query sets.

The geometry score predicts agreement with outcomes under policies outside the query set. On the 14 test games that count lives, we run each game's PPO policy from the 600 states. A distance is scored by how well it ranks the difference in lives lost between two states, after their difference in episode progress is removed. Figure 5 plots this agreement for every distance against its geometry. The native distance of the encoder reaches 0.273, close to the measured distance. No other native distance exceeds 0.20, and each is below the encoder with an interval that excludes zero. A head fit on the measurement raises every pretrained model above 0.20. Across the 43 distances the geometry score predicts the agreement at a correlation of 0.93, and at least 0.87 at every earlier stage of PPO training. The relation is unchanged without the encoder and holds among the native distances and among the heads separately. Appendix J adds the reward outcome and a second query set built from PPO checkpoints.

The form of the objective matters as much as the supervision. In Figure 6(a), regressing the fingerprint coordinates alone recovers less of the geometry than aligning distances to the measured ones. The combined objective of Equation 7 gives up a little of the aggregate score for robustness.



Representation	Native	Head
QueryPrint encoder	**0.273**	
VC-1 ViT-L/16	0.158	**0.288**
VIP ResNet-50	0.126	0.278
CMD	0.079	0.271
R3M ResNet-50	0.143	0.267
CURL	0.068	0.261
V-JEPA 2-g (clip)	0.144	0.252
DINOv2-B/14	0.161	0.249
V-JEPA 2-L (clip)	0.149	0.246
SigLIP 2 So400m/16	0.195	0.225
Random encoder	0.141	0.131

Figure 5: **Outcome agreement against geometry.** Each point is a distance, and the dashed line is the measured distance. Agreement is with lives lost under PPO policies. The table lists native and head agreement for the eight best heads, the best native distance, and a random encoder.

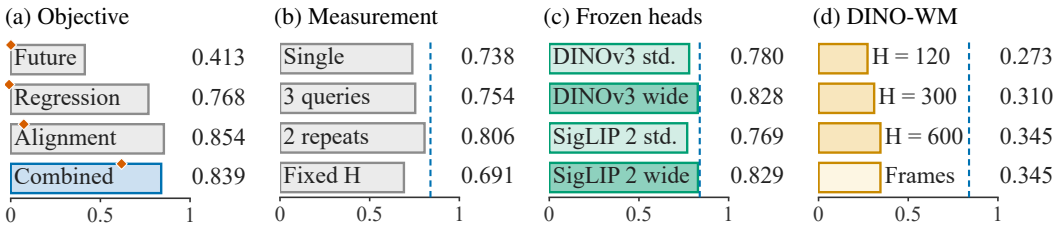


Figure 6: **Design study.** Geometry of every variant. (a) Forms of the objective, with a diamond marking the worst game. (b) Ablations of the measurement. (c) Behavioral heads on frozen features, standard (std.) and capacity-matched (wide). (d) Fingerprints imagined inside DINO-WM by horizon H, and at $H = 600$ against a frame-change target (Frames). Dashed lines mark the encoder.

Under alignment alone, *Freeway* drops to 0.073, and with both terms no game falls below 0.618. An encoder trained only to predict the same replayed futures (Future), at the same measurement cost, exposes about half of the geometry in its distance. A behavioral head on its embedding recovers most of it. The embedding holds the geometry, and the target exposes it in the distance.

Each part of the measurement matters. The ablations in Figure 6(b) use the best single measurement instead of the three combined ones, half the queries, half the repeats, or a single measurement with its horizon fixed at 600 steps. Every change lowers the score. Halving the repeats lowers it the least, and fixing the horizon lowers it the most, to 0.691, because the best horizon differs from game to game (§ 5.2).

Measured supervision also transfers the geometry to pretrained features. We give DINOv3 and SigLIP 2 a trained projection and a wider head, together close to our encoder in parameter count. In Figure 6(c), these heads, unlike the standard ones, come within about 0.01 of the encoder on the training states. On the additional states the encoder outperforms DINOv3 and is on par with SigLIP 2. The full tables of the three studies are provided in Appendix K.

Fingerprints imagined inside DINO-WM recover less than half of the measured geometry. We train DINO-WM (Zhou et al., 2025) per game on 2,048 stored trajectory segments, replay the test query policies inside it from the 600 states, and compute fingerprints from the change of its latent state. Figure 6(d) shows that at this budget their scores remain below 0.5 at every horizon, reaching 0.345 at the longest. The score rises with the horizon, indicating that the imagined futures carry part of the structure. The Frames row scores the same fingerprints against a target built from frame changes alone, the signal the model observes, and the score is unchanged. The gap therefore does not stem from signals the model cannot observe. Our encoder outperforms the world model on every game. Appendix L gives the training details of the world model.

6 Discussion and Conclusion

QueryPrint measures behavioral distance from counterfactual futures instead of learning it from trajectories, tests its validity, and uses it to supervise and to evaluate representations. On 55 Atari games, the baseline representations do not expose the measured geometry in their own distances, although frozen features hold it and a small head trained on the measurement recovers most of it. A compact encoder trained from pixels on the measurement gives the distance from observations alone and matches or exceeds these heads, under queries not used in training and on states it was not fit to. The measured distance also predicts where PPO policies outside the query set lose a life, an ability that only distances supervised by the measurement inherit. The same measurement tests a learned world model on the behavioral structure it implies rather than on the frames it renders.

Limitations. The measurement requires a resettable environment and emulator steps that the learned objectives do not use. The measured distance is relative to the query set, although two disjoint halves of it define closely agreeing geometries and the distance predicts outcomes under PPO policies outside it. The measurement is defined on states, whereas every representation sees observations, and the encoder and every head score lower beyond the training states (§ 5.3). Improving this generalization across episodes remains an open problem. We leave to future work whether closeness under the measured geometry helps control and whether one encoder can serve several games.

Reproducibility statement

Sections 4 and 5 and Appendices A to N specify the measurement, the validity tests, the training procedure, and every comparison. The distance targets, state snapshots, encoder embeddings, per-game scores, and the code that fits and scores representations are provided as supplementary material at `https://anonymous.4open.science/r/queryprint/`.

AI use statement

We used generative AI tools to polish the writing and to assist with code development. All AI-assisted text and code were reviewed, tested, and verified by the authors, who take full responsibility for the content of this paper.

References

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=qda7-sVg84`.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021b.

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. *Advances in neural information processing systems*, 32, 2019.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47: 253–279, 2013.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. Mico: Improved representations via sampling-based state similarity for markov decision processes. *Advances in Neural Information Processing Systems*, 34:30113–30126, 2021.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 162–169, Arlington, Virginia, USA, 2004. AUAI Press. ISBN 0974903906.

Norm Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden. Methods for computing state similarity in markov decision processes. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pp. 174–181, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

Scott Fujimoto, Pierluca D'Oro, Amy Zhang, Yuandong Tian, and Michael Rabbat. Towards general-purpose model-free reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 83509–83552, 2025.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial intelligence*, 147(1-2):163–223, 2003.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL http://dx.doi.org/10.1038/s41586-025-08744-2.

Michael R James and Satinder Singh. Learning and discovery of predictive state representations in dynamical systems with reset. In *Proceedings of the twenty-first international conference on Machine learning*, 2004.

Mete Kemertas and Tristan Aumentado-Armstrong. Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems*, 34:4764–4777, 2021.

Kim G Larsen and Arne Skou. Bisimulation through probabilistic testing. *Information and computation*, 94(1):1–28, 1991.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pp. 5639–5650. PMLR, 2020.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006. URL http://rbr.cs.umass.edu/aimath06/proceedings/P21.pdf.

Ziyan Luo, Tianwei Ni, Pierre-Luc Bacon, Doina Precup, and Xujie Si. Understanding the effectiveness of learning behavioral metrics in deep reinforcement learning. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=QqvrpgNtho.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YJ7o2wetJ2.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=TpU38jbKIJ.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xQiYCmDrjp.

Vivek Myers, Bill Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 19654–19679. Curran Associates, Inc., 2025. doi: 10.52202/085713-0661. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1c5956164472d6d8123d574aa75cd063-Paper-Conference.pdf.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=tGbpgz6yOrI.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Seohong Park, Deepinder Mann, and Sergey Levine. Dual goal representations. In *International Conference on Learning Representations*, volume 2026, pp. 153307–153324, 2026.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of machine learning research*, 22(268):1–8, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yutaka Shimizu and Masayoshi Tomizuka. Bisimulation metric for model predictive control. In *International Conference on Learning Representations*, volume 2025, pp. 93639–93660, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-2FCwDKRREu.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=D5RNACOZEI.

APPENDIX

A Query policies and measurement details . . . 14
B Algorithm . . . 15
C Implementation and budget of the learned objectives . . . 16
D Pretrained models and input frames . . . 17
E Statistics, hyperparameters, and compute . . . 19
F Results by game split . . . 20
G Validity statistics and reliability . . . 21
H Update and state budgets . . . 23
I Additional states and evaluation sets . . . 24
J Policies outside the query set . . . 25
K Objective, head, and measurement ablations . . . 28
L World model . . . 29
M Fitting the PSE metric offline . . . 30
N Results by game . . . 31

A QUERY POLICIES AND MEASUREMENT DETAILS

Table 2: **The six query policies.**

Query policy	Set	Actions
Uniform random	training	a uniformly random action resampled every block
No-op	training	the no-op action repeated
Horizontal	training	right and left in alternation, each held for 15 blocks
Held random	test	one uniformly sampled action held for 15 blocks
Fire and wait	test	one block of fire followed by three blocks of no-op
Vertical	test	up and down in alternation, each held for 15 blocks

Query policies. Six procedural policies act in the full 18-action space in blocks of four emulator steps (Table 2). The training query policies are uniform random actions resampled every block, the no-op action repeated, and alternating right and left held for 15 blocks each. The test query policies are a uniformly sampled action held for 15 blocks, one block of fire followed by three blocks of no-op, and alternating up and down held for 15 blocks each. The two sets share no policy but match in kind, since each has a random, a waiting, and a directional policy. Their agreement therefore shows how far the geometry depends on the particular policies. Appendix J measures a query set of a different kind. The query policies need no skill, and the measurement requires diversity between queries rather than competence, as the action-dependence test rewards queries whose futures differ.

States and signals. States are collected from episodes played from reset by the six query policies in turn, one per episode, and 120 snapshots are saved at each of five stages of episode progress. Every saved state is measured under all six queries, regardless of the query policy that produced its episode. The emulator reports the reward, termination, and lives at every step. The change signals are a fixed random projection to 16 dimensions of the change in the 128-byte emulator memory, the in-game events, and a fixed random projection to 16 dimensions of the change in the frame. Both projections are random so that the measurement needs no knowledge of what a byte or a pixel means, and both have 16 dimensions so that the memory and frame families yield fingerprints of the same size. The memory signal is computed from the same state that the emulator saves and restores, so it is available wherever the measurement applies, including the simulator state of a physics engine.

What the tests remove. The tests decide which part of a projected change counts. A coordinate that moves at the same rate from every state, such as a frame counter, is constant across states after standardization and drops out of every distance. A coordinate whose change does not depend on the query, such as scripted motion, falls in the shared part of Equation 3, which the shuffle test removes, and the non-degeneracy rule rejects a measurement it dominates. A coordinate that changes at random, such as a random number generator, averages out over the repeats and lowers reliability.

Grid, thresholds, and standardization. Discounted accumulation uses $\gamma \in \{0.9, 0.99, 0.999\}$ over $H \in \{120, 300, 600\}$, which gives 27 candidate measurements per game, or 18 in the 31 games without in-game event annotations. The validity thresholds are $\tau_{\text{shuffle}} = 0.90$, $\tau_{\min} = 0.70$, $\tau_{\text{rel}} = 0.97$, and $\tau_{\text{weak}} = 0.70$, and $k = 3$ candidates form the target. The outcome summaries appended to the accumulation are the discounted sum of the reward signs, an indicator of termination, the number of lives lost within the horizon, and an indicator that the rollout reached the horizon. With the four repeat statistics, these summaries occupy 16 of the 80 coordinates of a memory or frame measurement and 16 of the 36 coordinates of an event measurement, a minority in every measurement. Standardization statistics are computed per coordinate from the training-query fingerprints, pooled over the training queries. The same scale applies to every query block of both query sets, so the test distance does not depend on the order of the test queries.

B ALGORITHM

Algorithm 1 Measuring behavioral distances with QueryPrint

Input: resettable environment, states $s_1, \ldots, s_N$, queries $\Pi = \Pi_{\text{train}} \cup \Pi_{\text{test}}$, horizon H, repeats R, candidate measurements $\mathcal{C}$, thresholds τ, number of combined measurements k

▷ Measure counterfactual futures
for every state s_i, query $\pi \in \Pi$, and repeat $r = 1, \ldots, R$ **do**
  Restore s_i, run π for H steps, and record the per-step signals $x_1, \ldots, x_H$
end for
▷ Build and test every candidate measurement
for every candidate $c \in \mathcal{C}$ **do**
  Summarize each recorded future by $u^c_\pi(s_i)$, the discounted accumulation $\sum_t \gamma_c^{t-1} x_t^c$ followed by the outcome summaries, and form the fingerprints $F^c(s_i)$ from the statistics over repeats
  Compute the distance matrices D^c_{train} from the training queries and D^c_{test} from the test queries
  Take the query-specific parts $\mathring{F}^c_\pi(s_i)$ of Equation 3, permute the query labels within every state, and compute the shuffle retention G_c as the 95th percentile over permutations
  Split the repeats in half and compute the reliability ρ^{SB}_c of D^c_{train}
end for
Keep the candidates that are action dependent, reliable, and not degenerate, $\mathcal{V} = \{c : G_c < \tau_{\text{shuffle}}$ and $\rho^{\text{SB}}_c \geq \tau_{\min}$ and not $(\rho^{\text{SB}}_c > \tau_{\text{rel}}$ and $G_c \geq \tau_{\text{weak}})\}$
Rank $\mathcal{V}$ by action dependence, concatenate the fingerprints of the k highest-ranked candidates into the target fingerprint F_{train}, compute the target D and, from the same k candidates under the test queries, D_{test}, or declare the environment unmeasurable if fewer than k candidates are valid
return target fingerprints F_{train}, target distance D, and test distances D_{test}

Algorithm 1 summarizes the measurement of § 4. The QueryPrint encoder of § 4.4 is trained on its outputs F_{train} and D with Equation 7, and no step of the measurement depends on it.

C IMPLEMENTATION AND BUDGET OF THE LEARNED OBJECTIVES

Table 3: **Learned objectives compared in this paper.** Target quantity and training data of each.

Objective	Approximates	Learned from	Futures per state
BS-MPC	bisimulation metric	on-policy transitions	one
DBC	bisimulation metric	on-policy transitions	one
MICo	bisimulation metric, sampled	on-policy transitions	one
PSE	policy similarity metric	trajectory pairs, reference policy	one
CMD	temporal distance	trajectories	one
TMD	temporal distance	trajectories	one
CURL	augmentation invariance	augmented observations	one
ST-DIM	temporal consistency	consecutive observations	one
MR.Q	value-linear latent	transitions, reward and dynamics	one
QueryPrint encoder	measured behavioral distance	replays from saved states	many

Table 3 lists what each objective approximates and the data it learns from. Each objective follows the implementation published by its authors and is trained on the same 600 training states with the same IMPALA encoder, under the validation rule of the QueryPrint encoder. The native distance is the metric that the objective itself defines for DBC, MICo, PSE, CMD, and TMD, and the mean absolute distance of the embedding otherwise. PSE fits its policy similarity metric on the stored transitions with a bounded fixed-point solver (Appendix M).

Table 4: **Selected update counts of the learned objectives.**

Objective	Minimum	Median	Upper quartile	Maximum	At 2,000
BS-MPC	100	1100	1600	2000	8 of 165
CMD	100	400	1000	2000	3 of 165
CURL	200	1900	2000	2000	61 of 165
DBC	200	1800	1900	2000	40 of 165
MICo	100	100	100	1800	0 of 165
MR.Q	200	900	1200	1900	0 of 165
ST-DIM	100	1900	2000	2000	53 of 165
TMD	600	1500	1800	2000	17 of 165
QueryPrint encoder	100	800	–	2000	–

Batches hold 128 states, and training runs for at most 2,000 updates. The number of updates is selected on a progress-stratified 20% of the states by the objective's own validation loss, and each encoder is then reinitialized and refit on all 600 states for the selected number. Because the comparison concerns representations, every method keeps its representation losses as published and omits the control losses that need an actor or a critic. MR.Q and BS-MPC unroll their dynamics over the one-step transitions that the stored data provide, and all target networks use the same exponential moving average rate. Table 4 reports the selected update counts, and Appendix H trains the objectives for fixed budgets of up to 8,000 updates.

D PRETRAINED MODELS AND INPUT FRAMES

All 12 models receive native 210×160 RGB frames with the transforms prescribed by their authors, which for R3M include a resize to 256 and a center crop to 224. Image models receive the current frame and video models receive a 16-frame clip, while the single-frame V-JEPA variant receives the current frame repeated 16 times. Features are the pooled or class-token outputs of each model, and the native distance of a frozen model is the cosine distance on these features.

Table 5: **Behavioral head with four input frames.** Image models on the current and on the four most recent frames, with the paired difference, on the 18 test games.

Model	Current frame head	Four-frame head	Paired difference
DINOv2-B/14	0.713	0.740	+0.027 [+0.00, +0.07]
DINOv2-L/14	0.714	0.740	+0.026 [−0.00, +0.08]
SigLIP 2 So400m/16	0.789	0.814	+0.024 [+0.00, +0.04]
VC-1 ViT-L/16	0.714	0.733	+0.019 [−0.00, +0.06]

Table 6: **Native distance with four input frames.** Native distance of the image models on the current and on the four most recent frames, with the paired difference, on the 18 test games.

Model	Current frame	Four frames	Paired difference
DINOv2-B/14	0.359	0.329	-0.030 [−0.08, +0.01]
DINOv2-L/14	0.357	0.349	-0.008 [−0.07, +0.03]
DINOv3-L/16	0.363	0.363	-0.001 [−0.05, +0.04]
SigLIP 2 So400m/16	0.365	0.366	+0.001 [−0.02, +0.02]

Table 7: **Native distance on standardized features.** Native distance of the frozen models before and after standardizing every coordinate on the 600 states of a game, with the paired difference.

Model	Native	Standardized	Paired difference
Frozen visual representations			
DINOv2-B/14	0.263 [0.20, 0.34]	0.200 [0.14, 0.27]	-0.063 [−0.091, −0.040]
DINOv2-L/14	0.263 [0.20, 0.34]	0.206 [0.15, 0.27]	-0.057 [−0.084, −0.035]
DINOv3-L/16	0.274 [0.22, 0.36]	0.221 [0.17, 0.28]	-0.053 [−0.089, −0.033]
SigLIP 2 So400m/16	0.262 [0.20, 0.34]	0.207 [0.16, 0.26]	-0.055 [−0.101, −0.024]
V-JEPA 2-L (frame)	0.253 [0.19, 0.33]	0.199 [0.14, 0.27]	-0.054 [−0.084, −0.030]
V-JEPA 2-L (clip)	0.247 [0.18, 0.33]	0.195 [0.14, 0.27]	-0.052 [−0.085, −0.018]
V-JEPA 2-g (clip)	0.235 [0.17, 0.31]	0.193 [0.14, 0.26]	-0.042 [−0.076, −0.010]
V-JEPA 2.1-L (clip)	0.234 [0.18, 0.30]	0.207 [0.16, 0.27]	-0.027 [−0.054, −0.001]
VLM2Vec-V2	0.261 [0.20, 0.34]	0.196 [0.15, 0.25]	-0.065 [−0.105, −0.035]
Frozen control-pretrained representations			
R3M ResNet-50	0.276 [0.21, 0.36]	0.226 [0.16, 0.31]	-0.050 [−0.071, −0.027]
VC-1 ViT-L/16	0.266 [0.20, 0.35]	0.211 [0.15, 0.28]	-0.055 [−0.079, −0.033]
VIP ResNet-50	0.292 [0.23, 0.37]	0.196 [0.14, 0.26]	-0.096 [−0.125, −0.070]

Table 5 reports the behavioral head of four image transformers on the same four frames that the QueryPrint encoder receives, on the 18 test games. Four frames add at most 0.027, and the best four-frame score, 0.814 for SigLIP 2 So400m/16, remains below the 0.832 of the encoder on the same games, so the main table reports the single-frame features. The native distance of four image models on four frames remains at its single-frame level (Table 6), so the single input frame does not explain the low native scores. Feature scaling is not the cause either, since standardizing every coordinate of the frozen features on the 600 states of a game lowers the native score of all 12 models (Table 7). The domain gap between Atari frames and the pretraining data does not account for them

alone. The learned objectives, trained on the frames of each game, score no higher (Table 1), and the behavioral head recovers most of the geometry from the same frozen features.

Table 8: **Common feature-space distances.** L1, L2, and cosine are computed on the same features or embeddings as the native distance. Rows marked with a dagger keep an author-defined metric as their native distance. The reference row is not ranked. Bold marks the best and underline the second.

Representation	Native	L1	L2	Cosine
Reference				
Random encoder	0.264 [0.19, 0.35]	0.264 [0.19, 0.35]	0.265 [0.20, 0.35]	0.264 [0.20, 0.35]
Learned objectives				
BS-MPC	0.263 [0.19, 0.36]	0.263 [0.19, 0.36]	0.264 [0.19, 0.36]	0.264 [0.19, 0.36]
CMD†	0.200 [0.14, 0.27]	0.199 [0.14, 0.26]	0.201 [0.15, 0.26]	0.192 [0.14, 0.25]
CURL	0.165 [0.12, 0.23]	0.165 [0.12, 0.23]	0.166 [0.12, 0.23]	0.166 [0.13, 0.23]
DBC†	0.211 [0.15, 0.30]	0.211 [0.15, 0.30]	0.211 [0.15, 0.30]	0.211 [0.15, 0.30]
MICo†	0.197 [0.14, 0.27]	<u>0.293</u> [0.22, 0.38]	**0.294** [0.22, 0.39]	**0.294** [0.22, 0.39]
MR.Q	0.248 [0.18, 0.34]	0.248 [0.18, 0.34]	0.248 [0.18, 0.34]	0.248 [0.18, 0.34]
PSE†	0.130 [0.09, 0.20]	0.167 [0.11, 0.24]	0.166 [0.12, 0.24]	0.134 [0.09, 0.20]
ST-DIM	0.152 [0.11, 0.21]	0.152 [0.11, 0.21]	0.152 [0.11, 0.21]	0.153 [0.11, 0.21]
TMD†	0.185 [0.12, 0.25]	0.199 [0.14, 0.27]	0.193 [0.13, 0.27]	0.193 [0.13, 0.27]
Frozen visual representations				
DINOv2-B/14	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]
DINOv2-L/14	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]	0.263 [0.20, 0.34]
DINOv3-L/16	0.274 [0.21, 0.36]	0.277 [0.22, 0.36]	0.276 [0.22, 0.36]	0.274 [0.22, 0.36]
SigLIP 2 So400m/16	0.262 [0.20, 0.34]	0.262 [0.20, 0.34]	0.262 [0.20, 0.34]	0.262 [0.20, 0.34]
V-JEPA 2-L (frame)	0.253 [0.19, 0.33]	0.252 [0.19, 0.33]	0.252 [0.19, 0.33]	0.253 [0.19, 0.33]
V-JEPA 2-L (clip)	0.247 [0.18, 0.33]	0.245 [0.17, 0.33]	0.245 [0.17, 0.33]	0.247 [0.18, 0.33]
V-JEPA 2-g (clip)	0.235 [0.17, 0.32]	0.231 [0.16, 0.31]	0.232 [0.17, 0.31]	0.235 [0.17, 0.31]
V-JEPA 2.1-L (clip)	0.234 [0.18, 0.31]	0.234 [0.17, 0.31]	0.232 [0.17, 0.30]	0.234 [0.18, 0.30]
VLM2Vec-V2	0.261 [0.20, 0.34]	0.261 [0.20, 0.34]	0.261 [0.20, 0.34]	0.261 [0.20, 0.34]
Frozen control-pretrained representations				
R3M ResNet-50	<u>0.276</u> [0.21, 0.36]	0.293 [0.23, 0.37]	0.275 [0.21, 0.35]	0.276 [0.21, 0.36]
VC-1 ViT-L/16	0.266 [0.20, 0.35]	0.267 [0.20, 0.35]	0.265 [0.20, 0.35]	0.266 [0.20, 0.35]
VIP ResNet-50	**0.292** [0.23, 0.37]	**0.294** [0.23, 0.37]	<u>0.292</u> [0.23, 0.36]	<u>0.292</u> [0.23, 0.37]

Table 8 scores every representation with the mean absolute, Euclidean, and cosine distances on the same features or embeddings, next to its native distance. The highest value under any of the three is 0.294, reached by MICo and VIP, and for 18 of the 22 representations all three lie within 0.01 of the native score. On their own embeddings, the common distances score above the author-defined metrics of MICo and PSE, by at most +0.097 and +0.037, and no choice of distance brings any representation near the level of a behavioral head.

Table 9: **Effective dimensionality of the trained embeddings.** Participation ratio and variance share of the first principal component, as medians over games and seeds, next to the native geometry.

Representation	Effective dimension	Top-component share	Native geometry IQM
BS-MPC	2.305	0.615	0.263
CMD	1.731	0.707	0.200
CURL	7.361	0.248	0.165
DBC	2.074	0.664	0.211
MICo	4.914	0.393	0.197
MR.Q	2.686	0.548	0.248
PSE	4.985	0.351	0.130
ST-DIM	4.101	0.364	0.152
TMD	9.560	0.205	0.185
Random encoder	2.792	0.553	0.264
QueryPrint encoder	9.830	0.222	0.839

Dimensional collapse does not explain the lowest native scores either. Table 9 gives the effective dimensionality of every trained embedding on the 600 states of a game, the participation ratio of its covariance eigenvalues. The three objectives with the lowest native scores, PSE, ST-DIM, and CURL, retain more effective dimensions than the random encoder. The bootstrapped objectives col-

lapse to between two and five dimensions and score at the level of the random encoder. Effective dimensionality therefore does not predict the order of the native scores. PSE, ST-DIM, and CURL fall below the random encoder for a different reason. A random encoder still orders states by appearance, which the futures follow in part, as the observation-aliasing reference of Table 1 shows. Each of these objectives instead arranges its embedding by its own criterion, which removes the appearance similarity that a random encoder retains without putting the futures in its place.

E STATISTICS, HYPERPARAMETERS, AND COMPUTE

Statistics. Interquartile means are computed over the pooled game-by-seed values. Confidence intervals resample games within each split, training and test, with 20,000 draws and keep the seeds of a resampled game together. Paired differences resample the same games for both representations. Resampling the 600 states within every game instead, with games and seeds fixed, gives intervals of $[0.836, 0.843]$ for the QueryPrint encoder and $[0.776, 0.786]$ for the DINOv3 head around the aggregate scores of Table 1. The reported intervals therefore reflect variation across games rather than across states.

Encoder. The QueryPrint encoder is an IMPALA-style network with three residual stages of 16, 32, and 32 channels on four 84×84 RGB frames, a 256-dimensional projection with dropout 0.1, a 128-dimensional trunk, and a head of two linear layers with a hidden width of 512 and a 256-dimensional output. Training uses AdamW with a learning rate of 3×10^{-4}, weight decay 10^{-4}, batches of 128 states, and at most 2,000 updates. Validation runs every 100 updates on a progress-stratified 20% of the states and ends training after 8 checks without improvement. Gradients are clipped at 1 for the encoder and the heads and at 10 for the learned objectives. The random projection P of Equation 7 is a fixed sign matrix scaled by the inverse square root of the fingerprint dimension.

Behavioral head. The behavioral head is a two-layer network with 197,632 parameters on a 128-dimensional input. For the frozen models, the input is a fixed random projection of the standardized features, which gives all 12 frozen models heads of one size regardless of the width of their features. For the learned objectives, the input is the embedding itself, at its own width when that width exceeds 128. The head is fit on the 600 training states with the objective and the validation rule of the encoder. It transforms the information in its input but cannot add to it. For the random encoder and MICo, the input offers little that carries over to the validation states. The head then trains for the full budget and still scores at the level of the native distance or slightly below it. Every learned objective and every head uses the same rule, with a progress-stratified 20% of the states as the validation subset.

Fingerprint prediction. Fingerprint prediction splits the episodes of the 600 states of a game into training, validation, and test episodes. It fits a multi-output ridge regression from the representation to the test-query fingerprint coordinates on the training episodes and selects its regularization on the validation episodes. The score is the Spearman correlation between predicted and measured values on the test episodes, averaged over the coordinates that vary on those episodes.

Compute. Measuring one game costs 8.64M emulator steps on CPU. On a consumer GPU, an encoder trains in about a minute, and a head is fit in about ten seconds. Experiments ran on consumer GPUs (RTX 4080, 4090, 3060 Ti, 2080 Ti) and partly on cloud GPUs for the world model.

F RESULTS BY GAME SPLIT

Table 10: **Behavioral geometry by game split.** Native distance and behavioral head on the 37 training and the 18 test games. Bold marks the best and underline the second.

Representation	37 training games		18 test games	
	native	head	native	head
Learned objectives				
BS-MPC	0.243 [0.17, 0.35]	0.204 [0.14, 0.31]	0.317 [0.17, 0.52]	0.304 [0.15, 0.51]
CMD	0.171 [0.11, 0.25]	0.498 [0.43, 0.57]	0.258 [0.17, 0.41]	0.540 [0.45, 0.65]
CURL	0.166 [0.12, 0.23]	0.722 [0.66, 0.77]	0.172 [0.09, 0.39]	0.760 [0.66, 0.81]
DBC	0.196 [0.13, 0.30]	0.314 [0.22, 0.41]	0.250 [0.15, 0.45]	0.366 [0.18, 0.57]
MICo	0.180 [0.12, 0.26]	0.184 [0.12, 0.27]	0.251 [0.13, 0.43]	0.296 [0.14, 0.50]
MR.Q	0.225 [0.16, 0.32]	0.299 [0.22, 0.39]	0.322 [0.16, 0.53]	0.418 [0.23, 0.61]
PSE	0.112 [0.07, 0.18]	0.552 [0.49, 0.60]	0.199 [0.10, 0.42]	0.595 [0.49, 0.68]
ST-DIM	0.140 [0.10, 0.20]	0.612 [0.54, 0.67]	0.204 [0.10, 0.38]	0.665 [0.58, 0.74]
TMD	0.160 [0.09, 0.25]	0.228 [0.17, 0.32]	0.235 [0.13, 0.34]	0.295 [0.18, 0.45]
Frozen visual representations				
DINOv2-B/14	0.219 [0.16, 0.31]	0.733 [0.68, 0.78]	0.359 [0.23, 0.48]	0.713 [0.62, 0.80]
DINOv2-L/14	0.219 [0.16, 0.31]	0.727 [0.67, 0.77]	0.357 [0.23, 0.48]	0.714 [0.61, 0.80]
DINOv3-L/16	0.247 [0.18, 0.33]	<u>0.775</u> [0.74, 0.81]	0.363 [0.22, 0.55]	0.791 [0.71, 0.84]
SigLIP 2 So400m/16	0.224 [0.17, 0.31]	0.760 [0.72, 0.80]	0.365 [0.22, 0.54]	0.789 [0.72, 0.84]
V-JEPA 2-L (frame)	0.232 [0.16, 0.33]	0.707 [0.65, 0.76]	0.303 [0.19, 0.44]	0.706 [0.60, 0.79]
V-JEPA 2-L (clip)	0.223 [0.15, 0.33]	0.733 [0.67, 0.78]	0.293 [0.18, 0.43]	0.740 [0.65, 0.81]
V-JEPA 2-g (clip)	0.210 [0.14, 0.31]	0.739 [0.69, 0.78]	0.289 [0.17, 0.43]	0.743 [0.66, 0.81]
V-JEPA 2.1-L (clip)	0.207 [0.14, 0.29]	0.758 [0.70, 0.80]	0.309 [0.18, 0.47]	<u>0.797</u> [0.72, 0.85]
VLM2Vec-V2	0.209 [0.16, 0.30]	0.759 [0.70, 0.81]	<u>0.376</u> [0.26, 0.52]	0.796 [0.73, 0.84]
Frozen control-pretrained representations				
R3M ResNet-50	0.244 [0.17, 0.35]	0.713 [0.66, 0.77]	0.339 [0.22, 0.46]	0.708 [0.61, 0.80]
VC-1 ViT-L/16	0.224 [0.16, 0.33]	0.736 [0.69, 0.78]	0.345 [0.24, 0.47]	0.714 [0.63, 0.80]
VIP ResNet-50	<u>0.278</u> [0.21, 0.36]	0.725 [0.67, 0.77]	0.329 [0.21, 0.46]	0.706 [0.62, 0.80]
QueryPrint encoder	**0.841** [0.82, 0.86]	**0.841** [0.82, 0.86]	**0.832** [0.79, 0.88]	**0.832** [0.79, 0.88]

Table 10 reports the two geometry columns of Table 1 on the 37 training and the 18 test games. The QueryPrint encoder reaches 0.841 $[0.819,\ 0.863]$ on the training games and 0.832 $[0.786,\ 0.878]$ on the test games. On both splits the QueryPrint encoder scores highest, the head of every frozen model exceeds 0.7, and CURL (0.722 and 0.760), ST-DIM (0.612 and 0.665), and CMD (0.498 and 0.540) keep their order among the learned objectives.

G VALIDITY STATISTICS AND RELIABILITY

Tests. The shuffle test subtracts the shared part of Equation 3, permutes the query labels of the recorded futures within each state, and recomputes the distance matrix. Retention is the Spearman correlation with the unshuffled matrix of the query-specific parts, reported as the 95th percentile over 200 permutations. The permutation is drawn independently within every state, so each draw relabels all 600 states at once rather than choosing among the six orderings of one state. Because no permutation changes the shared part, keeping it would raise the retention of every measurement whether or not its futures depend on the queries. Reliability and the target distance are computed on the full fingerprint. The split-half test partitions the four repeats into two pairs in two ways, builds a distance matrix from each half, and reports the Spearman-Brown correction of the mean correlation. The reliability floor rejects measurements with reliability below 0.70, and the non-degeneracy rule rejects measurements with reliability above 0.97 and retention at or above 0.70.

Table 11: **Validity statistics of every game.** Retention and reliability on the test queries.

Game	Retention	Reliability	Split	Game	Retention	Reliability	Split
Alien	0.51	0.93	train	Krull	0.81	0.95	train
Amidar	0.50	0.92	train	Kung Fu Master	0.48	0.94	train
Assault	0.34	0.95	train	Montezuma's Revenge	0.19	0.93	train
Asterix	0.52	0.94	train	Ms. Pac-Man	0.48	0.92	test
Asteroids	0.53	0.93	test	Name This Game	0.30	0.97	train
Atlantis	0.63	0.95	train	Phoenix	0.50	0.91	train
Bank Heist	0.16	0.91	train	Pitfall	0.36	0.96	test
Battle Zone	0.45	0.90	test	Pong	-0.02	0.91	train
Beam Rider	0.85	0.94	train	Private Eye	0.26	0.94	train
Berzerk	0.21	0.90	test	Qbert	0.38	0.94	test
Bowling	0.09	0.92	test	Riverraid	0.07	0.97	test
Boxing	0.16	0.96	test	Robotank	0.52	0.96	train
Breakout	0.04	0.92	test	Seaquest	0.44	0.90	test
Centipede	0.34	0.95	train	Skiing	0.55	0.97	test
Chopper Command	0.29	0.98	train	Solaris	0.26	0.96	train
Crazy Climber	0.34	0.94	train	Space Invaders	0.48	0.94	test
Defender	0.30	0.95	train	Star Gunner	0.70	0.97	train
Demon Attack	0.31	0.88	train	Surround	0.69	0.98	train
Double Dunk	0.17	0.97	train	Tennis	0.05	0.95	train
Enduro	0.43	0.95	train	Time Pilot	0.38	0.96	train
Fishing Derby	0.19	0.94	train	Tutankham	0.21	0.97	train
Freeway	0.74	0.69	test	Up N Down	0.51	0.92	train
Frostbite	0.34	0.94	test	Venture	0.50	0.85	test
Gopher	0.03	0.89	train	Video Pinball	0.38	0.95	test
Hero	0.16	0.97	train	Wizard Of Wor	0.76	0.95	train
Ice Hockey	0.41	0.94	train	Yars Revenge	0.28	0.97	test
Jamesbond	0.47	0.89	train	Zaxxon	0.52	0.95	train
Kangaroo	0.37	0.90	train				

Statistics of every game. Table 11 lists retention and reliability of every target on the test queries. The reliability floor applies to the training-query statistics at selection, and the test-query reliabilities are computed only afterwards. A selected target can therefore fall marginally below the floor on the test queries, as one game, *Freeway*, does. Its scores are compared with its own test-query reliability, and without this game the aggregate score of the QueryPrint encoder is 0.840. Every selected measurement remains valid when retention is computed between independent halves of the repeats, so that the permuted and the unpermuted distances share no measurement noise.

Sensitivity of the selection rule. Compared with selection without a reliability floor, the floor of 0.70 changes the target of 1 game, and floors of 0.80, 0.85, and 0.90 change 6, 7, and 10 of the 55 games. Every game still has at least three valid candidates under each of these floors. The target of every game is unchanged under a shuffle threshold of 0.85 or 0.95, a reliability threshold of 0.99, or a weak-retention threshold of 0.60 or 0.80. A reliability threshold of 0.95 changes the target of 1 game and the score of the encoder to 0.836. An alternative rule ranks the valid candidates by the product of reliability and action dependence instead of action dependence alone. Its targets agree with the selected ones at a median of 1.000, and the score of the encoder is 0.836. On the 18 test games alone, the main pattern holds. Every native distance other than the encoder's is below 0.4, the heads of the frozen models recover most of the geometry, and the encoder scores highest (Appendix F).

Table 12: **Control geometries built from parts of the fingerprint.** Spearman correlation with the test-query target, as the median, minimum, and maximum over games. Outcome statistics are the lives, termination, reward, and horizon coordinates recorded under the queries.

Control geometry	Median	Minimum	Maximum
All outcome statistics, per query	0.503	0.243	0.927
Outcome statistics, query averaged	0.496	0.243	0.927
Query-averaged outcomes plus progress	0.483	0.114	0.903
Top-member outcomes, query averaged	0.430	0.169	0.917

Progress, episodes, and outcome statistics. The target distance correlates with the progress distance between two states, the absolute difference of their step counts within the episode, at a median of 0.062 over games. The correlation is highest in *Ms. Pac-Man*, at 0.500, and below 0.33 in every other game. Its correlation with episode identity has a median of 0.101. Progress and episode membership explain little of the geometry, at most 25% of its rank variance in any game. Table 12 scores distances built from parts of the fingerprint against the target. The outcome statistics alone, with or without progress, reach a median of only about 0.5 and exceed 0.7 in 7 games dominated by events or survival.

Reliability. Fresh query rollouts from the stored 600 states of *Montezuma's Revenge* and *Amidar* reproduce the stored test-query distances with Spearman correlations of 0.920 and 0.919, close to the split-half reliabilities of those games. The split-half statistic is therefore a faithful estimate of test-retest reliability.

Aliasing. The observation-aliasing reference of Table 1 gives every state the mean fingerprint of the states that share its four-frame history and scores the result against the test queries. The reference is the score of a fingerprint that depends on the observation alone and reaches 0.964, above the split-half reliability because it carries no measurement noise. In the median game, 412 of the 600 states have distinct four-frame histories, against 48 in *Breakout*. The reference falls below 1 only because of the states that share a history. It bounds what any function of the observation can score on these states and says nothing about how easily that function is learned.

Emulator memory. A further control asks whether the target is a simple function of the emulator state. On the 18 test games we score the distance between the memory of two states, the 128 bytes saved with each state, as we would score a representation. With every byte standardized over the 600 states, the mean absolute difference reaches 0.488 $[0.366, 0.609]$. The same distance reaches 0.478 on the raw bytes, and the share of bytes that differ reaches 0.478. The standardized distance lies above the native distance of every frozen model on these games, which is at most 0.376, and below the head of every frozen model, which is at least 0.706. The QueryPrint encoder, which sees only pixels, outperforms it in all 18 games, by +0.344 $[+0.215, +0.476]$. The memory distance is lowest in the 6 games whose highest-ranked measurement is of the memory family, at a median of 0.339, against 0.609 and 0.555 for the event and frame families. The target is measured in part from how the memory changes under the queries, and the content of the memory predicts it only in part.

H UPDATE AND STATE BUDGETS

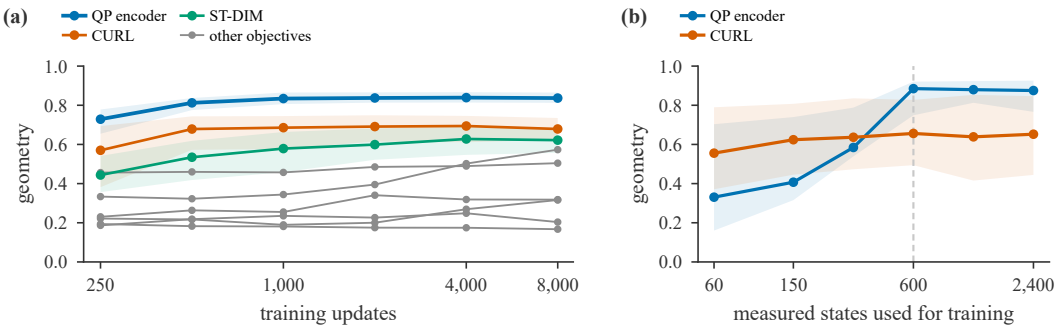


Figure 7: **Update and state budgets.** (a) Geometry against the number of training updates without early stopping on 20 games with one seed, with the learned objectives scored through the behavioral head and shaded 95% confidence intervals over games. (b) Geometry on the 600 states of five training games against the number of measured states used for training.

Table 13: **Geometry under fixed update budgets.** Bold marks the best value in a column and underline the second best.

Objective / method	250	1,000	2,000	4,000	8,000
BS-MPC	0.222 [0.13, 0.36]	0.189 [0.11, 0.33]	0.200 [0.12, 0.34]	0.269 [0.18, 0.39]	0.316 [0.22, 0.42]
CMD	0.456 [0.37, 0.54]	0.457 [0.38, 0.54]	0.485 [0.41, 0.55]	0.489 [0.44, 0.56]	0.504 [0.46, 0.58]
CURL	<u>0.570</u> [0.38, 0.69]	<u>0.686</u> [0.58, 0.74]	<u>0.691</u> [0.61, 0.75]	<u>0.694</u> [0.61, 0.75]	<u>0.679</u> [0.62, 0.74]
DBC	0.230 [0.11, 0.38]	0.255 [0.12, 0.43]	0.341 [0.19, 0.48]	0.319 [0.17, 0.46]	0.318 [0.19, 0.44]
MICo	0.187 [0.10, 0.32]	0.235 [0.14, 0.36]	0.226 [0.14, 0.36]	0.249 [0.17, 0.37]	0.204 [0.13, 0.33]
MR.Q	0.334 [0.18, 0.50]	0.344 [0.18, 0.52]	0.395 [0.20, 0.59]	0.502 [0.31, 0.65]	0.573 [0.40, 0.69]
ST-DIM	0.444 [0.36, 0.54]	0.579 [0.46, 0.66]	0.599 [0.52, 0.68]	0.628 [0.55, 0.69]	0.622 [0.56, 0.68]
TMD	0.194 [0.13, 0.31]	0.181 [0.12, 0.30]	0.175 [0.12, 0.29]	0.175 [0.12, 0.29]	0.167 [0.11, 0.29]
QP encoder	**0.729** [0.65, 0.78]	**0.834** [0.81, 0.87]	**0.837** [0.81, 0.87]	**0.839** [0.81, 0.87]	**0.837** [0.81, 0.86]

For Figure 7(a) and Table 13, the encoder and every learned objective except PSE are trained for fixed numbers of updates without early stopping on a random subset of 20 games. The encoder saturates by 1,000 updates at 0.834 and remains there at 8,000. MR.Q gains the most from a larger budget, from 0.395 at 2,000 updates to 0.573 at 8,000, and no objective exceeds 0.694 at any budget.

Table 14: **Geometry against the number of measured states used for training.**

Method	60	150	300	600	1,200	2,400
QP encoder	0.331 [0.16, 0.70]	0.407 [0.32, 0.74]	0.584 [0.54, 0.79]	0.885 [0.75, 0.92]	0.880 [0.81, 0.92]	0.875 [0.77, 0.93]
CURL	0.555 [0.37, 0.79]	0.624 [0.45, 0.81]	0.637 [0.47, 0.84]	0.656 [0.49, 0.83]	0.639 [0.42, 0.85]	0.652 [0.44, 0.85]

For Figure 7(b) and Table 14, we train the QueryPrint encoder and the CURL encoder on varying numbers of states from five training games and score every model on the same 600 states. For the QueryPrint encoder these are measured states. CURL needs no measurement to train its encoder, but its head is fit on all 600 measured states at every budget, so the two methods use the same measured states only at 600 states. Below 600 states, some of the scored states were not used to train the encoder, so these points measure generalization to other states rather than how well the training states are fit. With all 600 states, the encoder reaches 0.743 on the states of a further episode (Appendix I). Beyond 600 states neither score rises, so 600 measured states per game are enough, and more states do not bring CURL closer.

I ADDITIONAL STATES AND EVALUATION SETS

Table 15: **Behavioral geometry on the additional states of the test games.** Bold marks the best and underline the second.

Representation	Native distance	Behavioral head
Learned objectives		
BS-MPC	0.237 [0.12, 0.45]	0.222 [0.11, 0.43]
CMD	0.183 [0.10, 0.34]	0.406 [0.23, 0.58]
CURL	0.214 [0.12, 0.38]	0.666 [0.50, 0.75]
DBC	0.208 [0.12, 0.41]	0.311 [0.13, 0.53]
MICo	0.124 [0.04, 0.32]	0.243 [0.10, 0.45]
MR.Q	0.263 [0.14, 0.48]	0.331 [0.16, 0.56]
PSE	0.198 [0.10, 0.36]	0.530 [0.35, 0.65]
ST-DIM	0.193 [0.11, 0.37]	0.580 [0.45, 0.68]
TMD	0.041 [−0.01, 0.11]	0.165 [0.06, 0.33]
Frozen visual representations		
DINOv2-B/14	0.215 [0.11, 0.36]	0.602 [0.41, 0.73]
DINOv2-L/14	0.219 [0.11, 0.37]	0.641 [0.46, 0.74]
DINOv3-L/16	0.273 [0.16, 0.46]	0.565 [0.45, 0.69]
SigLIP 2 So400m/16	0.284 [0.17, 0.46]	0.688 [0.57, 0.77]
V-JEPA 2-L (frame)	0.178 [0.08, 0.38]	0.599 [0.41, 0.70]
V-JEPA 2-L (clip)	0.201 [0.09, 0.39]	0.603 [0.45, 0.73]
V-JEPA 2-g (clip)	0.227 [0.09, 0.42]	0.654 [0.50, 0.75]
V-JEPA 2.1-L (clip)	0.254 [0.16, 0.43]	<u>0.721</u> [0.63, 0.79]
VLM2Vec-V2	<u>0.346</u> [0.22, 0.49]	0.653 [0.56, 0.73]
Frozen control-pretrained representations		
R3M ResNet-50	0.226 [0.11, 0.40]	0.548 [0.40, 0.66]
VC-1 ViT-L/16	0.224 [0.14, 0.37]	0.643 [0.45, 0.75]
VIP ResNet-50	0.246 [0.10, 0.44]	0.536 [0.39, 0.66]
Random encoder	0.199 [0.09, 0.43]	0.160 [0.08, 0.39]
QueryPrint encoder	**0.743** [0.65, 0.82]	**0.743** [0.65, 0.82]

Additional states of the test games. Each test game has 100 stored states that no representation was fit to, drawn from one further episode of the same collection. Table 15 scores every representation on them with the native distance and with the behavioral head fit on the 600 training states. In paired differences over the 18 games, the QueryPrint encoder leads the DINOv3 head by +0.178 $[+0.066, +0.294]$, the SigLIP 2 head by +0.055 $[+0.009, +0.118]$, and the V-JEPA 2.1 head, the closest, by +0.022 $[-0.010, +0.058]$.

Table 16: **Evaluation sets of the paper.**

Evaluation set	Games	Rule
Main table	55	the 600 measured states of every game that passes the validity tests
Training split	37	games used to develop the method
Test split	18	the remaining games
Update budget curve	20	random draw of 10 training and 10 test games
State budget curve	5	random draw among training games with enough stored states
Additional states	18	100 stored states per test game that no representation was fit to
Expert query set	6	training games measured a second time with a query set of expert checkpoints
Outcome agreement	18	test games under PPO policies, lives lost on the 14 that count lives
Emulator memory control	18	test games, distance between the memory bytes of two states
World model	55	all games, three seeds each

Evaluation sets. Table 16 lists every set of games used in the paper and the rule that defines it.

J POLICIES OUTSIDE THE QUERY SET

Table 17: **Expert query set on six training games.** Agreement between the procedural and the expert query sets, reliability of each, and scores against the expert test-query distances.

Game	Valid	Agreement		Reliability		Score, expert target	
	candidates	train	test	expert	procedural	encoder	random
Demon Attack	19	0.655	0.561	0.936	0.877	0.606	0.119
Hero	17	0.826	0.778	0.973	0.967	0.818	0.837
Montezuma's Revenge	27	0.795	0.719	0.869	0.932	0.736	0.403
Pong	10	0.910	0.749	0.974	0.907	0.740	0.357
Private Eye	16	0.768	0.672	0.971	0.936	0.635	0.046
Tennis	27	0.880	0.863	0.987	0.950	0.882	0.673

Agreement with an expert query set. On six training games we measure the same 600 states a second time with a query set of six policies, five of them from PPO training. These are the checkpoints at 5, 20, 50, and 100 percent of training with deterministic actions, the final checkpoint with sampled actions, and uniform random actions. The 5 and 50 percent checkpoints and the random policy serve as training queries and the rest as test queries, and the target is selected with the rule of § 4.3. Between 10 and 27 candidates pass the validity tests under this query set. The test-query distances of the two query sets agree at a mean of 0.724 $[0.647, 0.793]$ over the six games, and the agreement of individual games ranges from 0.561 to 0.863. The training-query distances agree at 0.806 $[0.735, 0.867]$, and the split-half reliabilities of the expert query set range from 0.869 to 0.987 (Table 17).

The QueryPrint encoder trained on the procedural target recovers the geometry of the expert query set at 0.736 $[0.660, 0.812]$, the level at which the two measurements agree with each other. The best native distance of a frozen representation, VIP, reaches 0.414, and the random encoder reaches 0.406. In *Hero* the expert geometry follows the layout of the screens, and the random encoder recovers it as well as the QueryPrint encoder does. *Hero* is one of the games of § 5.3 in which a head on a random encoder scores high.

Outcomes under PPO policies. Agreement between two query sets still compares one measurement with another. A more direct test leaves the measurement altogether and asks what the distance says about a policy that was never a query. If two states are close, such a policy should fare similarly from both. We test this on the 18 test games with a PPO agent (Schulman et al., 2017) trained on each game for 10M steps (Raffin et al., 2021), under the sticky actions and the full action space of the measurement. Neither the target nor any representation or head has seen its outcomes. Only PSE has seen the policy itself, since its metric is fitted from the policy's action distributions (Appendix M), so it is left out of this test.

The outcome we score is the loss of a life, the event of Figure 1. The loss cannot be undone, and a trained policy acts to avoid it. Where it occurs under such a policy reflects the situation of the state more than the quality of the policy. This outcome is defined in the 14 test games that count lives. The reward collected over the same rollout is a second outcome, scored on the 16 test games in which it differs between states. In *Venture* the policy collects none, and in *Skiing* reward is a time penalty that accrues at the same rate from every state.

From each of the 600 states the policy acts with sampled actions for 40 decisions of four emulator steps. To bound the cost of the rollouts, 100 states stratified by episode progress are rolled out three times and the remaining 500 are rolled out once. Each of the 100 is paired with each of the 500, which gives 50,000 pairs per game. A distance is scored by the Spearman correlation between its values and the absolute difference in outcome over these pairs. The absolute difference in episode progress is removed from both by partial correlation, because two states far apart in an episode differ in distance and in outcome for reasons unrelated to behavior. Each representation is scored with its distance averaged over seeds, a behavioral head is fit on the training-query target of the same 600 states, and scores are aggregated over games as in Appendix E.

Table 18: **Outcome agreement under the final PPO policy.** Bold marks the best value in a column and underline the second best, and the random encoder and the measured distances are not ranked.

Representation	Lives lost, 14 games		Reward, 16 games	
	native distance	behavioral head	native distance	behavioral head
Learned objectives				
BS-MPC	0.126 [0.07, 0.21]	0.146 [0.07, 0.23]	0.106 [0.05, 0.21]	0.061 [0.02, 0.19]
CMD	0.079 [0.03, 0.15]	0.271 [0.21, 0.35]	0.086 [0.03, 0.20]	0.132 [0.06, 0.24]
CURL	0.068 [0.03, 0.17]	0.261 [0.19, 0.35]	0.065 [0.03, 0.13]	0.136 [0.07, 0.23]
DBC	0.099 [0.03, 0.21]	0.158 [0.08, 0.26]	0.114 [0.05, 0.20]	0.105 [0.04, 0.21]
MICo	0.094 [0.03, 0.18]	0.153 [0.06, 0.24]	0.069 [0.03, 0.19]	0.089 [0.03, 0.19]
MR.Q	0.130 [0.05, 0.24]	0.165 [0.09, 0.26]	0.101 [0.05, 0.19]	0.093 [0.05, 0.19]
ST-DIM	0.051 [0.01, 0.13]	0.204 [0.16, 0.28]	0.043 [0.01, 0.12]	0.112 [0.05, 0.20]
TMD	0.079 [0.02, 0.16]	0.133 [0.05, 0.23]	<u>0.116</u> [0.05, 0.19]	0.084 [0.04, 0.18]
Frozen visual representations				
DINOv2-B/14	0.161 [0.08, 0.25]	0.249 [0.17, 0.34]	0.108 [0.06, 0.18]	0.127 [0.06, 0.22]
DINOv2-L/14	0.157 [0.07, 0.23]	0.228 [0.16, 0.32]	0.093 [0.05, 0.18]	0.122 [0.06, 0.21]
DINOv3-L/16	0.187 [0.11, 0.25]	0.225 [0.16, 0.32]	0.108 [0.05, 0.20]	0.105 [0.03, 0.20]
SigLIP 2 So400m/16	<u>0.195</u> [0.11, 0.26]	0.225 [0.16, 0.33]	0.101 [0.04, 0.21]	0.106 [0.04, 0.20]
V-JEPA 2-L (frame)	0.104 [0.03, 0.20]	0.230 [0.15, 0.33]	0.113 [0.06, 0.19]	0.126 [0.05, 0.23]
V-JEPA 2-L (clip)	0.149 [0.07, 0.23]	0.246 [0.18, 0.35]	0.098 [0.04, 0.19]	0.139 [0.06, 0.23]
V-JEPA 2-g (clip)	0.144 [0.06, 0.24]	0.252 [0.18, 0.36]	0.099 [0.04, 0.18]	<u>0.144</u> [0.06, 0.24]
V-JEPA 2.1-L (clip)	0.139 [0.06, 0.20]	0.229 [0.15, 0.34]	0.104 [0.03, 0.20]	0.104 [0.04, 0.19]
VLM2Vec-V2	0.185 [0.10, 0.26]	0.218 [0.17, 0.32]	0.094 [0.04, 0.20]	0.086 [0.02, 0.19]
Frozen control-pretrained representations				
R3M ResNet-50	0.143 [0.07, 0.23]	0.267 [0.21, 0.36]	0.093 [0.04, 0.19]	0.142 [0.06, 0.23]
VC-1 ViT-L/16	0.158 [0.08, 0.24]	**0.288** [0.22, 0.37]	0.105 [0.06, 0.18]	**0.145** [0.08, 0.24]
VIP ResNet-50	0.126 [0.05, 0.23]	<u>0.278</u> [0.22, 0.36]	0.084 [0.05, 0.17]	0.140 [0.07, 0.24]
Random encoder	0.141 [0.06, 0.24]	0.131 [0.07, 0.21]	0.078 [0.03, 0.16]	0.072 [0.01, 0.18]
QueryPrint encoder	**0.273** [0.22, 0.35]	0.273 [0.22, 0.35]	**0.139** [0.06, 0.23]	0.139 [0.06, 0.23]

Measured distances, not ranked. Test-query distance 0.265 [0.23, 0.34] on lives lost and 0.148 [0.07, 0.23] on reward, training-query distance 0.278 [0.23, 0.36] on lives lost and 0.130 [0.06, 0.21] on reward.

The measured distance predicts where the PPO policy loses a life, at 0.265, and the QueryPrint encoder exposes this in its own distance, at 0.273 (Table 18). This is the level of the training-query distance it is fit to, 0.278, so the encoder preserves what its target predicts about a policy that took no part in its training. No other native distance exceeds 0.20, and all 21 fall below the encoder with a paired interval that excludes zero. The encoder is higher than each of them in at least 11 of the 14 games. The gain is large where survival depends on small objects, as in § 5.3. In *Battle Zone* and *Asteroids* no native distance exceeds 0.12, whereas the encoder reaches 0.27 and 0.26. The head brings every frozen model from between 0.104 and 0.195 to between 0.218 and 0.288. The VC-1 head scores highest and the encoder is close behind, at a paired difference of -0.015 $[-0.058, +0.031]$. The head also brings the contrastive objectives CURL, CMD, and ST-DIM to 0.261, 0.271, and 0.204, whereas the bootstrapped metric objectives and the random encoder gain little, as in Table 1.

Table 19: **Outcome agreement by PPO checkpoint, for lives lost.** Group columns average the distances of a group within every game, and the last column is the Pearson correlation between geometry and agreement across distances.

PPO checkpoint	Measured distance	QP encoder	Frozen models		Learned objectives		Correlation with geometry
			native	head	native	head	
5% of training	0.260	0.258	0.159	0.220	0.074	0.161	0.87
20% of training	0.272	0.287	0.156	0.247	0.081	0.183	0.94
50% of training	0.278	0.278	0.165	0.237	0.087	0.167	0.91
100% of training	0.265	0.273	0.164	0.232	0.093	0.188	0.93

The correspondence holds across all distances in Table 18. The geometry of a distance is its Spearman correlation with the measured test-query distance on the same pairs. It predicts the outcome agreement of the distance at a Pearson correlation of 0.93 over the 43 distances (Figure 5), and between 0.87 and 0.94 at every PPO checkpoint from 5 to 100 percent of training (Table 19). For the final policy the correlation is 0.93 without the encoder, 0.82 among the native distances alone, and

0.92 among the heads alone. The geometry score thus orders distances by how well they anticipate the outcomes of a policy outside the query set, regardless of its competence.

Without the control for progress the ordering is unchanged, and the encoder reaches 0.283 against 0.197 for the highest native distance. Reward accrues from most states, in 70% of the rollouts of the median game against 17% for the loss of a life, and separates states less sharply. The measured distance itself reaches 0.148 on reward, and the ordering repeats at smaller margins. The encoder reaches 0.139, and no native distance exceeds 0.116.

K OBJECTIVE, HEAD, AND MEASUREMENT ABLATIONS

Table 20: **Objective ablation.** The minimum is the lowest score over games and seeds, and games below 0.3 counts the games in which the lowest-scoring seed falls below 0.3.

Objective	Geometry	Minimum	Games below 0.3
Future prediction, no geometric term	0.413	0.001	17
Regression only	0.768	−0.009	3
Alignment only	0.854	0.073	1
Regression and alignment	0.839	0.618	0

Objective. Table 20 compares the two terms of Equation 7 on the 55 games with the same network, data, and rule, together with an encoder trained on the same counterfactual futures without any geometric term. In this future-prediction encoder, a predictor with the architecture of the behavioral head receives the embedding and a learned embedding of the training query identity. The predictor is trained with a Huber loss to predict the standardized repeat means of every candidate measurement under that query. This encoder sees no pairwise distance, no selected target, and no test query, and it follows the training rule of the QueryPrint encoder. Alignment alone gives the highest aggregate score but falls to 0.073 on *Freeway*, and the combined objective is the only one whose worst game remains above 0.3. The future-prediction encoder reaches 0.413 $[0.372,\ 0.464]$ with its own distance and 0.815 $[0.790,\ 0.837]$ through a head.

Table 21: **Capacity-matched heads on frozen features.** Trained parameters counts what is fit on the measurement, the head for frozen features and the whole network for the QueryPrint encoder.

Representation	Head	Trained parameters	Geometry	Additional states	Prediction
DINOv3-L/16	standard	197,632	0.780	0.565	0.303
DINOv3-L/16	trained projection	1,313,536	0.828	0.638	0.549
SigLIP 2 So400m/16	standard	197,632	0.769	0.688	0.282
SigLIP 2 So400m/16	trained projection	1,379,072	0.829	0.715	0.538
QueryPrint encoder	none	1,321,680	0.839	0.743	0.564

Head capacity. Table 21 compares two heads. The behavioral head of Table 1 has 197,632 parameters on a fixed random projection to 128 dimensions. The wider head adds a trained linear projection from the frozen feature to 512 dimensions and uses a wider network, for a total close to the 1,321,680 parameters of the QueryPrint encoder. On the training states the wider head closes the gap, and in fingerprint prediction it closes most of it. On the additional states of the test games the QueryPrint encoder outperforms DINOv3 by +0.104 $[+0.015,\ +0.204]$ and is on par with SigLIP 2, at +0.028 $[-0.026,\ +0.094]$.

Table 22: **Measurement ablations.** Score of each variant against the full target, with the paired difference to the full method.

Variant	Geometry	Δ
Single measurement	0.738	−0.101 $[-0.137,\ -0.073]$
3 queries	0.754	−0.085 $[-0.115,\ -0.060]$
2 repeats	0.806	−0.033 $[-0.050,\ -0.022]$
Fixed horizon	0.691	−0.147 $[-0.207,\ -0.097]$

Measurement. Table 22 reports the measurement ablations discussed in § 5.4. Each variant trains the QueryPrint encoder on a target with one component of the measurement changed and is scored against the test-query distances of the full target. The three-query variant keeps one training and two test queries, so its target comes from a single training query.

L WORLD MODEL

Model and training. We adapt DINO-WM (Zhou et al., 2025) to Atari, following the implementation published by its authors. The model predicts frozen DINOv3 ViT-L/16 patch features from a history of three frames spaced four emulator steps apart. The features are 196 tokens of 1,024 dimensions on the native frame, resized to a shorter edge of 224 pixels and center-cropped to a square. The predictor is a six-layer causal ViT with 16 attention heads and an action embedding of 10 dimensions concatenated to every patch token, trained with a mean squared error on the next-step patch features. Each four-step block of discrete actions is encoded as four concatenated one-hot vectors, and the proprioceptive input of the original model, which Atari lacks, is removed. Each predictor is trained with AdamW at a learning rate of 5×10^{-4} and batch size 32 for 100 epochs, with three seeds per game. Training and validation use 2,048 and 256 trajectory segments, drawn from stored trajectories of the game under an episode-disjoint split, and the checkpoint with the lowest validation error is kept.

Imagined fingerprints. Fingerprints are computed inside the model from the three most recent of the four frames stored with each of the 600 states, which initialize its history. The action sequences of the test queries and their repeats are replayed inside the model for 30, 75, and 150 predictor steps, which correspond to 120, 300, and 600 emulator steps. Each imagined future is summarized by the discounted accumulation of the signed spatial mean and of the spatial mean absolute value of the predicted change in patch features, with discounts 0.9, 0.99, and 0.999. Mean, standard deviation, and 10th and 90th percentiles over repeats form the fingerprint, which is standardized as in Equation 2. The mean absolute distance between fingerprints is scored against the measured test-query distances of the same states. The fingerprint is built from the change of the predicted features along an imagined rollout, not from a distance between features, so the native score of DINOv3 does not bound it.

Table 23: **Geometry of the fingerprints imagined inside DINO-WM.** Scores by horizon against the measured target and against the highest-ranked frame-change measurement of each game.

Horizon	Measured target	Frame-change target
120	0.273 [0.22, 0.34]	0.281 [0.22, 0.35]
300	0.310 [0.26, 0.37]	0.314 [0.26, 0.38]
600	0.345 [0.30, 0.41]	0.345 [0.30, 0.41]

Frame-change target. We also score the same imagined fingerprints against the highest-ranked frame-change measurement of each game, the signal family the model observes. The world model then reaches 0.345 $[0.295, 0.406]$ at the 600-step horizon against 0.345 $[0.299, 0.408]$ for the full target, a paired difference of 0.000 $[-0.045, +0.042]$. At the shorter horizons the scores are 0.281 and 0.314. The gap to the measurement therefore does not stem from the signal families the model cannot observe. Table 23 lists both scores at every horizon.

M FITTING THE PSE METRIC OFFLINE

Of the learned objectives, policy similarity embeddings (PSE) (Agarwal et al., 2021a) have the target nearest to the measured distance, since both compare two states by what the same policy does from them. The policy similarity metric of PSE is the fixed point of $d(x, y) = \mathrm{TV}\big(\pi(x), \pi(y)\big) + \gamma\, d(x', y')$, where TV is the total variation distance between the action distributions of a reference policy π at the two states, x' and y' are the states that follow them, and γ is a discount. The implementation published by its authors computes the metric by dynamic programming along paired trajectories of π from the two states. The stored transitions of the offline single-game setting are one-step tuples and provide no such trajectories, so we fit the metric differently, as described below.

The reference policy is the final checkpoint of the PPO agent of Appendix J, evaluated on the stored states and on their successors, since the optimal policy assumed by the original method is unavailable. We fit the metric on the 600 states by fitted-value iteration with a small critic. The iteration uses a discount of 0.99, an iterate threshold of 0.02 and a Bellman threshold of 0.15 on the median of the last five outer iterations, and at most 450 iterations. Because the metric is bounded by $1/(1-\gamma) = 100$, the bootstrapped values are clipped to that bound at every iteration. All 55 games pass the convergence thresholds. The PSE encoder is then trained with the contrastive loss of the original method on the fitted metric under the same rule as every other objective. Its native distance reaches 0.130, and its behavioral head reaches 0.566 (Table 1).

N RESULTS BY GAME

Table 24: **Target of every game.** The top member is the highest-ranked of the three concatenated measurements, given by its signal family, its horizon H in emulator steps, and its discount γ. Retention and reliability are on the test queries, and the query agreement is the Spearman correlation between the training-query and the test-query distances.

Game	Top target member			Shuffle	Split-half	Query
	Family	H	γ	retention	reliability	agreement
Training games						
Alien	memory	600	0.9	0.508	0.925	0.869
Amidar	frame	300	0.999	0.498	0.924	0.879
Assault	memory	600	0.9	0.338	0.950	0.904
Asterix	memory	600	0.999	0.524	0.937	0.838
Atlantis	memory	120	0.99	0.629	0.954	0.910
Bank Heist	frame	120	0.99	0.162	0.907	0.774
Beam Rider	memory	600	0.999	0.853	0.936	0.835
Centipede	memory	120	0.9	0.342	0.952	0.904
Chopper Command	frame	600	0.99	0.294	0.979	0.934
Crazy Climber	memory	300	0.999	0.338	0.940	0.867
Defender	frame	120	0.999	0.299	0.949	0.889
Demon Attack	memory	600	0.99	0.311	0.877	0.780
Double Dunk	memory	120	0.99	0.169	0.969	0.962
Enduro	memory	120	0.99	0.431	0.952	0.917
Fishing Derby	memory	120	0.999	0.193	0.941	0.941
Gopher	memory	120	0.9	0.034	0.894	0.892
Hero	frame	120	0.9	0.165	0.967	0.953
Ice Hockey	memory	600	0.999	0.413	0.936	0.845
Jamesbond	memory	600	0.9	0.470	0.892	0.814
Kangaroo	memory	600	0.9	0.368	0.904	0.794
Krull	memory	300	0.999	0.809	0.946	0.901
Kung Fu Master	frame	600	0.999	0.482	0.944	0.885
Montezuma's Revenge	memory	600	0.99	0.193	0.932	0.905
Name This Game	frame	300	0.999	0.304	0.967	0.939
Phoenix	memory	120	0.9	0.499	0.907	0.864
Pong	memory	120	0.9	−0.020	0.907	0.926
Private Eye	memory	120	0.999	0.262	0.936	0.842
Robotank	memory	120	0.99	0.524	0.956	0.905
Solaris	memory	300	0.999	0.262	0.961	0.897
Star Gunner	frame	120	0.999	0.700	0.968	0.915
Surround	memory	600	0.999	0.688	0.975	0.942
Tennis	frame	300	0.999	0.047	0.950	0.943
Time Pilot	memory	120	0.9	0.380	0.964	0.941
Tutankham	memory	600	0.999	0.212	0.971	0.946
Up N Down	memory	600	0.999	0.514	0.923	0.859
Wizard Of Wor	memory	120	0.999	0.755	0.955	0.882
Zaxxon	memory	600	0.999	0.520	0.955	0.935
Test games						
Asteroids	frame	600	0.999	0.532	0.933	0.836
Battle Zone	memory	300	0.9	0.446	0.902	0.808
Berzerk	event	600	0.99	0.210	0.902	0.776
Bowling	frame	300	0.999	0.086	0.924	0.935
Boxing	memory	300	0.999	0.162	0.962	0.899
Breakout	frame	600	0.999	0.044	0.924	0.904
Freeway	event	600	0.999	0.741	0.694	0.628
Frostbite	event	600	0.999	0.340	0.940	0.908
Ms. Pac-Man	frame	600	0.999	0.476	0.915	0.845
Pitfall	event	120	0.999	0.361	0.961	0.933
Qbert	memory	600	0.9	0.381	0.938	0.937
Riverraid	event	120	0.999	0.070	0.968	0.945
Seaquest	event	300	0.999	0.438	0.898	0.795
Skiing	memory	120	0.99	0.546	0.973	0.916
Space Invaders	memory	120	0.99	0.477	0.941	0.905
Venture	event	120	0.9	0.502	0.852	0.835
Video Pinball	memory	120	0.99	0.378	0.948	0.916
Yars Revenge	frame	600	0.99	0.276	0.968	0.954

Table 24 lists the target of every game with its validity statistics on the test queries and the agreement of its two query sets. Table 25 gives the geometry of every game for the QueryPrint encoder under its native distance and for the behavioral heads of the random encoder, of DINOv3, and of the nine learned objectives, averaged over three seeds. Games are grouped by split and ordered alphabetically within a group.

Table 25: **Geometry by game.** The QueryPrint encoder under its native distance and the behavioral head on the other representations, averaged over three seeds. Bold marks the best value in a row and underline the second best, and the random encoder is not ranked.

Game	QP encoder	Behavioral head										
	native	Random	DINOv3	BS-MPC	CMD	CURL	DBC	MICo	MR.Q	PSE	ST-DIM	TMD
Training games												
Alien	**0.855**	0.126	0.683	0.160	0.336	0.641	0.251	0.185	0.106	0.327	0.548	0.142
Amidar	**0.870**	0.291	0.856	0.447	0.604	0.854	0.429	0.367	0.361	0.682	0.764	0.118
Assault	**0.872**	0.165	0.445	0.066	0.372	0.371	0.339	0.079	0.336	0.488	0.345	0.222
Asterix	**0.805**	0.088	0.748	0.048	0.313	0.712	0.390	0.076	0.066	0.565	0.667	0.258
Atlantis	0.759	0.293	**0.763**	0.195	0.470	0.656	0.536	0.213	0.302	0.529	0.655	0.138
Bank Heist	0.839	0.831	**0.862**	0.814	0.790	0.843	0.790	0.821	0.838	0.855	0.851	0.649
Beam Rider	**0.751**	0.282	0.710	0.320	0.405	0.586	0.274	0.268	0.324	0.520	0.513	0.192
Centipede	**0.873**	0.140	0.843	0.107	0.291	0.810	0.458	0.182	0.347	0.610	0.745	0.163
Chopper Command	**0.927**	0.713	0.911	0.620	0.781	0.869	0.708	0.592	0.551	0.675	0.786	0.627
Crazy Climber	**0.821**	0.126	0.786	0.093	0.508	0.670	0.102	0.166	0.106	0.484	0.536	0.369
Defender	**0.812**	0.171	0.727	0.022	0.351	0.499	0.214	0.073	0.088	0.383	0.257	0.261
Demon Attack	**0.780**	0.123	0.770	0.207	0.360	0.727	0.146	0.088	0.189	0.584	0.261	0.301
Double Dunk	**0.846**	0.521	0.832	0.553	0.702	0.808	0.301	0.588	0.350	0.725	0.733	0.608
Enduro	**0.840**	0.017	0.647	0.007	0.089	0.220	0.052	−0.002	0.063	0.173	0.204	0.157
Fishing Derby	**0.913**	0.015	0.890	0.035	0.199	0.790	0.028	0.080	0.029	0.654	0.727	0.080
Gopher	**0.775**	0.448	0.754	0.546	0.606	0.668	0.529	0.531	0.545	0.626	0.663	0.383
Hero	0.716	0.694	**0.803**	0.721	0.722	0.753	0.704	0.689	0.706	0.718	0.743	0.673
Ice Hockey	**0.729**	0.060	0.573	0.171	0.228	0.280	0.166	0.230	0.163	0.243	0.256	0.176
Jamesbond	**0.795**	0.024	0.699	0.101	0.430	0.695	0.070	0.060	0.328	0.541	0.641	0.199
Kangaroo	**0.773**	0.024	0.736	0.137	0.532	0.612	0.201	0.112	0.179	0.157	0.567	0.168
Krull	**0.875**	0.350	0.812	0.264	0.572	0.794	0.574	0.364	0.548	0.518	0.539	0.112
Kung Fu Master	**0.855**	0.301	0.807	0.246	0.521	0.779	0.374	0.290	0.294	0.583	0.598	0.076
Montezuma's Revenge	**0.919**	0.052	0.913	0.595	0.753	0.841	0.562	0.205	0.385	0.783	0.801	0.607
Name This Game	**0.926**	0.479	0.841	0.377	0.557	0.822	0.635	0.387	0.395	0.526	0.713	0.312
Phoenix	**0.825**	0.091	0.733	0.151	0.405	0.630	0.371	0.055	0.452	0.530	0.615	0.139
Pong	**0.855**	0.202	0.693	0.226	0.589	0.659	0.123	0.127	0.130	0.401	0.614	0.437
Private Eye	**0.795**	0.012	0.633	0.049	0.389	0.511	0.038	0.049	0.162	0.320	0.440	0.221
Robotank	**0.786**	0.067	0.654	0.012	0.246	0.464	0.119	−0.016	0.263	0.311	0.284	0.091
Solaris	0.688	0.521	**0.706**	0.498	0.599	0.608	0.514	0.506	0.580	0.590	0.599	0.365
Star Gunner	**0.888**	−0.032	0.863	0.256	0.679	0.777	0.513	0.105	0.125	0.530	0.591	0.337
Surround	**0.917**	0.324	0.895	0.304	0.655	0.899	0.317	0.268	0.273	0.692	0.798	0.096
Tennis	**0.930**	0.461	0.882	0.665	0.690	0.874	0.624	0.575	0.636	0.819	0.834	0.578
Time Pilot	**0.856**	0.153	0.621	0.162	0.376	0.498	0.243	0.117	0.259	0.285	0.429	0.123
Tutankham	**0.887**	0.482	0.863	0.077	0.817	0.859	0.025	0.053	0.646	0.635	0.755	0.688
Up N Down	**0.825**	0.082	0.762	0.046	0.454	0.785	0.072	0.056	0.125	0.328	0.347	0.168
Wizard Of Wor	**0.846**	−0.016	0.806	−0.018	0.539	0.781	0.248	0.056	−0.010	0.625	0.545	0.140
Zaxxon	**0.871**	0.439	0.739	0.249	0.528	0.754	0.150	0.237	0.570	0.599	0.580	0.124
Test games												
Asteroids	**0.791**	0.127	0.687	0.129	0.562	0.741	0.449	0.269	0.696	0.635	0.670	0.133
Battle Zone	**0.772**	0.074	0.537	0.081	0.344	0.421	0.066	0.076	0.071	0.214	0.342	0.156
Berzerk	0.784	0.551	**0.790**	0.480	0.699	0.788	0.579	0.475	0.693	0.749	0.759	0.464
Bowling	**0.726**	0.361	0.586	0.373	0.592	0.542	0.390	0.381	0.383	0.408	0.460	0.479
Boxing	**0.835**	0.089	0.607	0.184	0.336	0.369	0.150	0.132	0.247	0.351	0.375	0.234
Breakout	0.840	0.784	0.876	0.799	0.803	**0.877**	0.789	0.759	0.773	0.796	0.809	0.759
Freeway	0.628	0.061	**0.630**	0.098	0.096	0.249	0.056	0.117	0.039	0.002	0.597	0.084
Frostbite	**0.896**	0.048	0.874	0.091	0.484	0.826	0.056	0.025	0.354	0.608	0.568	0.292
Ms. Pac-Man	**0.838**	0.129	0.806	0.565	0.588	0.739	0.661	0.525	0.574	0.479	0.641	0.100
Pitfall	**0.909**	0.549	0.780	0.580	0.645	0.692	0.563	0.606	0.555	0.620	0.668	0.422
Qbert	**0.935**	0.223	0.932	0.214	0.484	0.902	0.214	0.190	0.208	0.550	0.810	0.314
Riverraid	0.913	0.784	**0.928**	0.731	0.837	0.880	0.816	0.754	0.804	0.891	0.872	0.738
Seaquest	0.797	0.235	**0.798**	0.274	0.441	0.760	0.179	0.190	0.198	0.667	0.701	0.311
Skiing	**0.886**	0.134	0.799	0.070	0.310	0.764	0.028	0.054	0.076	0.661	0.768	0.061
Space Invaders	**0.882**	0.064	0.814	0.052	0.440	0.783	0.353	0.069	0.285	0.435	0.592	0.152
Venture	**0.734**	0.062	0.730	0.103	0.468	0.726	0.126	0.100	0.091	0.613	0.526	0.116
Video Pinball	0.753	0.515	**0.821**	0.684	0.705	0.758	0.673	0.679	0.677	0.653	0.668	0.571
Yars Revenge	0.928	0.703	**0.935**	0.730	0.827	0.917	0.809	0.709	0.853	0.825	0.873	0.642