REWEIGHT, DO NOT RE-CALIBRATE

Reusing the same encoder–predictor pair (world model) when the policy or query changes.

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model calibrated under one policy is queried under another once a planner selects actions from its predictions. We ask whether the same frozen encoder-predictor can be reused under a changed policy or a changed query law without re-calibration on target labels. Conformal off-policy prediction marginalizes the action and predicts a return, which makes the importance weight a function of the outcome and forces an outcome model, rejection subsampling, or a trajectory product. We condition on the action and predict the next state, which makes the weight measurable before the outcome and removes all three; the reduction holds at one step, and the occupancy factor returns over a horizon. With a planner enacting the new policy on RoboCasa, exact propensity weighting of the source calibration set reaches 0.913 coverage with no new observations, against 0.910 after recalibrating on 75 labelled target episodes per task (model seed 271828, temperature 3); recalibration re-collects 300 labelled episodes under each new policy, while the weighted construction reuses one 600-episode source set across all nine target policies. We prove that the target-to-source likelihood ratio is measurable before the outcome and factors into state-occupancy and action-propensity ratios, and combine weighted conformal prediction with the exact or an estimated ratio. Under a query change, coverage at nominal 0.90 moves from 0.902 to 0.550 and to 0.909 with learned weights on 192 Franka transitions, and from 0.734 to 0.903 on 5,100 FurnitureBench episodes, matching oracle weighting; with V-JEPA-2-AC in place of DINOv2 on the Franka data, 0.561 and 0.917. Under the planner-enacted policy change, coverage moves from 0.750 to 0.912 with each radius finite, a paired gain of 16.2 points [8.5, 23.7], replicating on two further model seeds at 0.690 to 0.877 and 0.735 to 0.892. For a learned ratio, the exact convex dual of the classifier-regret constraint within each quantile bin of the posterior gives coverage lower bounds of 0.863 and 0.802 on the two query-change corpora, against observed 0.909 and 0.903. Finite-support identities of the argument are machine-checked in Lean 4.[1]

1 INTRODUCTION

We use world models to predict the consequences of candidate actions in latent imagination, model-predictive planning, physical robot learning, and predictive representation learning (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025; Wu et al., 2023; Zhou et al., 2025; Terver et al., 2025). Planners already select actions from these predictions in the physical world (Chua et al., 2018; Hansen et al., 2024), and the corpora that support this are physical-robot corpora (Khazatsky et al., 2024; Heo et al., 2023): an over-confident prediction set becomes a collision rather than a bad number. A capable agent cannot avoid carrying such a model (Richens et al., 2025), so the operative quantity is how much to trust it, and for robotics in particular a world model needs a calibrated estimate of the certainty of its own predictions (Karcini et al., 2026). Among the roles a predictive dynamics model can occupy—deploy-time planner, learned simulator, data generator, policy representation, safety monitor, evaluation tool (Do Robots Need World Models? Workshop Organizers, 2026)—we study the last two. In the evaluator role the bar is not realism but whether the rollout remains reliable enough

[1] Our proofs close all goals in the Lean development and use no project-specific axioms. Source: anonymized commit-pinned Lean project.

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 ranking, selection, and rejection (Hou et al., 2026); in deployment the requirement is reliability without failure data or task-specific supervision (Agia, 2026). Performance degrades under perturbation even within the same environment (Jangir et al., 2026), and a change of policy is such a perturbation on the query side. When a planner selects actions from a world model's predictions, the target policy generates a different state–action distribution from the behavior policy that supplied calibration data. We can keep the transition mechanism and predictor fixed while changing the law under which we evaluate prediction errors. If we retain the source calibration quantile, we calibrate under the wrong measure.

The coverage of the prediction set varies along the variable whose law we change. On the Franka corpus (Section 6), under the source law at $\lambda = 1.5$ and the 0.90 level, the DINOv2 prediction set covers 1.000, 1.000 and 0.731 of the transitions in the lower, middle and upper thirds of the tilt statistic g, and the V-JEPA-2-AC set covers 1.000, 0.972 and 0.741; these are population values over the 192 transitions. The upper third holds the largest issued actions. When our change leaves the law of the next state given the query unchanged (A1), the score $R = r_m(X, Y)$ and the likelihood ratio $w = \mathrm{d}P_t / \mathrm{d}P_b$ of the target law to the source law (Section 2) satisfy, for any threshold q,

$$P_t(R \le q) - P_b(R \le q) = \mathrm{Cov}_{P_b}\big(w,\ \mathbf{1}\{R \le q\}\big),$$

because $\mathbb{E}_{P_b}[w] = 1$. When we change the law through a variable V, so that w is a function of V, the covariance equals $\mathrm{Cov}_{P_b}\big(w(V), P_b(R \le q \mid V)\big)$: unweighted coverage stays at its source level under all such changes exactly when $P_b(R \le q \mid V)$ does not vary with V, and it falls when we move weight toward values of V where $P_b(R \le q \mid V)$ is low. On the Franka corpus $V = g$, and when we tilt the query law toward large actions we put its weight on the upper third, where coverage is lowest; the old calibration set holds too few of the transitions where the model errs most, and unweighted DINOv2 coverage, averaged over 50 replicates, falls to 0.550. We reweight the calibration set by the likelihood ratio so that it matches the target law on those transitions.

We use conformal prediction for finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018). Weighted conformal prediction extends this construction to covariate shift when the likelihood ratio is available (Tibshirani et al., 2019). The shift-robust conformal literature otherwise either adapts online to a shift it does not name (Gibbs & Candès, 2021), bounds how far coverage falls when the data depart from exchangeability, without identifying the departure (Barber et al., 2023), or protects against a ball of shifts rather than the one a policy induces (Cauchois et al., 2024; Zhou et al., 2024). Zhao et al. (2025) also treat a changing policy, updating the calibration set online as target outcomes arrive during interaction; we hold the calibration set fixed and change only the measure under which it is read, so no target outcome is observed at any point.

Conformal prediction has been carried to off-policy problems before, and the estimand there is a return: Taufiq et al. (2022) write the joint under a policy with the action marginalized out, and state that the policy change shifts the conditional law of the outcome given the context rather than the covariate law. The weight is then a function of the outcome, and each route around that fact has a drawback. Estimating it directly needs an outcome model and an integration over candidate outcomes (Taufiq et al., 2022; Foffano et al., 2023); the subsampling route recovers an outcome-free weight but retains a selected subsample whose size falls exponentially in the number of decision stages (Zhang et al., 2023); the trajectory product inherits the horizon. We take the action into the covariate and predict the next state instead of the return. The transition mechanism is invariant to the policy, so the change of policy is a covariate shift on the state–action marginal, and the weight is measurable before the outcome (P1). We therefore need no outcome model, candidate grid, or discarded calibration point. A policy intervention determines the ratio through the two policies and the state occupancy, and that ratio must be identified from quantities available before the predicted outcome. Off-policy learning separates action-propensity and state-occupancy correction (Liu et al., 2018; Nachum et al., 2019; Jiang & Li, 2016); covariate-shift learning estimates density ratios directly or through domain classification (Shimodaira, 2000; Sugiyama et al., 2007; Bickel et al., 2009; Menon & Ong, 2016); propensity-weighted conformal inference quantifies coverage loss from approximate weights (Lei & Candès, 2021).

We ask one question: under an intervention on action selection, with the conditional transition mechanism invariant, which change of measure transports marginal coverage from source to target, how much coverage do we lose when we estimate it, and when do unlabeled target covariates recover the guarantee? We identify the likelihood ratio from pre-outcome covariates, factor it into occupancy and propensity, transfer weighted conformal coverage through total variation, and bound ratio error by

domain-classifier regret. We use Pearl's intervention semantics and the separation between population variation and invariant mechanisms in transportability analysis to state the premise (Pearl, 2009; Pearl & Bareinboim, 2014).

We test the construction with two kinds of evidence (Section 5). Under a *query change*, the target action law is declared and evaluation reweights recorded executed state–action pairs from physical robots: 192 Franka transitions on two frozen visual encoders and 5,100 FurnitureBench episodes (Heo et al., 2023). Under a *policy change*, a frozen latent world-model planner enacts the target policy on RoboCasa and outcomes are drawn fresh from restored simulator states, which recorded data cannot supply; the same source calibration set is compared with recalibration on labelled target episodes. The conformal construction is fixed throughout.

2 We identify the target change of measure

Let $X \in \mathcal{X}$ contain information available when we issue a prediction and let $Y \in \mathcal{Y}$ denote the predicted outcome. For one-step world-model prediction, we take $X = (S, A)$ and Y as the next observation or latent target. We fix a world model m, choose a measurable nonconformity score r_m, and set $R = r_m(X, Y)$.

Assumption 1 (Covariate shift)**.** We assume $P_{\mathrm{t},X} \ll P_{\mathrm{b},X}$ and a common Markov kernel K such that

$$P_b(\mathrm{d}x, \mathrm{d}y) = P_{\mathrm{b},X}(\mathrm{d}x)K(\mathrm{d}y \mid x), \qquad P_t(\mathrm{d}x, \mathrm{d}y) = P_{\mathrm{t},X}(\mathrm{d}x)K(\mathrm{d}y \mid x). \tag{A1}$$

Proposition 1 (Pre-outcome identification)**.** *Under A1,*

$$\frac{\mathrm{d}P_t}{\mathrm{d}P_b}(x, y) = \frac{\mathrm{d}P_{\mathrm{t},X}}{\mathrm{d}P_{\mathrm{b},X}}(x) \qquad P_b\text{-a.e.} \tag{P1}$$

Because the action is a covariate rather than a marginalized nuisance, and the predicted quantity is the next state rather than a return, the ratio carries no dependence on y. Off-policy conformal prediction is obliged to work with a weight that does depend on it (Taufiq et al., 2022; Foffano et al., 2023; Zhang et al., 2023). We use P1 to separate prediction from calibration. We fit m to approximate K; the deployment policy determines the distribution of the query X.

For a policy-controlled transition, write

$$P_b(\mathrm{d}s, \mathrm{d}a, \mathrm{d}y) = d_b(\mathrm{d}s)\,\pi_b(\mathrm{d}a \mid s)\,K(\mathrm{d}y \mid s, a), \tag{1}$$
$$P_t(\mathrm{d}s, \mathrm{d}a, \mathrm{d}y) = d_t(\mathrm{d}s)\,\pi_t(\mathrm{d}a \mid s)\,K(\mathrm{d}y \mid s, a). \tag{2}$$

Proposition 2 (Policy-intervention ratio)**.** *If $d_t \ll d_b$ and $\pi_t(\cdot \mid s) \ll \pi_b(\cdot \mid s)$ for d_t-almost all s, then*

$$w(s, a) = \frac{\mathrm{d}P_{\mathrm{t},X}}{\mathrm{d}P_{\mathrm{b},X}}(s, a) = \underbrace{\frac{\mathrm{d}d_t}{\mathrm{d}d_b}(s)}_{\textit{state occupancy}} \underbrace{\frac{\mathrm{d}\pi_t(\cdot \mid s)}{\mathrm{d}\pi_b(\cdot \mid s)}(a)}_{\textit{action propensity}}. \tag{P2}$$

When we intervene at one step from a common state population, $d_t = d_b$ and the propensity ratio is the whole correction. At later query times the robot reaches different states under the new policy, so both factors remain. When we declare both policies and the transition mechanism stays the same, the two policies and the initial-state law determine the occupancy ratio, and we do not estimate it from visited states. Deployed decision systems log the probability of each chosen action at decision time (Agarwal et al., 2016; Swaminathan & Joachims, 2015; Gauci et al., 2018; Saito et al., 2021), and we read the source density from such a log. When a density is unavailable, we estimate the product of the two factors from source and target covariates (Section 4).

3 We transport marginal coverage

Let $(X_i, Y_i)_{i=1}^n \overset{\text{iid}}{\sim} P_b$ be a calibration sample and let $R_i = r_m(X_i, Y_i)$. For a nonnegative weight v and test covariate x, weighted conformal prediction places mass proportional to $v(X_i)$ at each calibration

score and the standard test-point mass at $+\infty$ (Tibshirani et al., 2019); the test point carries normalized mass $p_{n+1}(x) = v(x)/\big(\sum_{i=1}^{n} v(X_i) + v(x)\big)$, so the radius is infinite exactly when $p_{n+1}(x) > \alpha$. We denote the resulting $(1-\alpha)$ quantile by $q_v(x)$ and return

$$C_v(x) = \{y : r_m(x, y) \le q_v(x)\}.$$

With the exact likelihood ratio $w = \mathrm{d}P_t / \mathrm{d}P_b$, this construction attains target-law marginal coverage at least $1-\alpha$. When the test-point mass carries more than α of the normalized weight, $q_v(x) = +\infty$ and the returned set is the whole outcome space; the rate of such sets is a measurement of the overlap between the two laws, and we report it alongside coverage. P3 is the coverage-loss bound for approximate weights of Lei & Candès (2021), restated for the policy-shift ratio of P2.

We next allow an estimated nonnegative ratio $\widehat{w}$, fitted independently of the calibration sample. Define

$$Z_{\widehat{w}} = \mathbb{E}_{P_b}[\widehat{w}(X)], \qquad \frac{\mathrm{d}\widehat{P}}{\mathrm{d}P_b} = \frac{\widehat{w}}{Z_{\widehat{w}}}.$$

Global rescaling leaves the weighted conformal quantile unchanged, so $\widehat{w}$ is the exact likelihood ratio for the surrogate target $\widehat{P}$.

Proposition 3 (Approximate-weight coverage transfer)**.** *Under A1, with $0 < Z_{\widehat{w}} < \infty$ and ratio fitting independent of calibration,*

$$\mathbb{P}_{P_b^n \otimes P_t}\{Y_{n+1} \in C_{\widehat{w}}(X_{n+1})\} \ge 1 - \alpha - \mathrm{TV}(P_t, \widehat{P}),$$

and

$$\mathrm{TV}(P_t, \widehat{P}) \le \frac{1}{2}\left(\mathbb{E}_{P_b}|w - \widehat{w}| + |Z_{\widehat{w}} - 1|\right).$$

[P3]

We use total variation because it gives the sharp distribution-only penalty for an arbitrary event. The second inequality bounds the penalty by the source-weighted L^1 error of the estimated likelihood ratio and the deviation of its normalizer from one. We keep efficiency separate. Concentrated weights reduce effective sample size and can enlarge prediction sets, but the validity penalty is the change of measure from P_t to $\widehat{P}$.

4 We learn the ratio from unlabeled covariates

When we use an estimated ratio, we lose at most $\frac{1}{2}\big(\mathbb{E}_{P_b}|w - \widehat{w}| + |Z_{\widehat{w}} - 1|\big)$ of coverage (P3). Work that transports coverage off-policy carries the same ratio-error term and leaves it as a nuisance to be assumed small (Taufiq et al., 2022; Zhang et al., 2023), and improving the estimate of that ratio is named the most significant open direction in that line (Foffano et al., 2023). We bound this term through the fitted classifier and its log-loss regret.

We pool source and target covariates with a domain label $D \sim \mathrm{Bern}(\rho)$, where

$$X \mid D = 0 \sim P_{\mathrm{b},X}, \qquad X \mid D = 1 \sim P_{\mathrm{t},X},$$

and let $M = (1-\rho)P_{\mathrm{b},X} + \rho P_{\mathrm{t},X}$. If $\eta(x) = \mathbb{P}(D = 1 \mid X = x)$, Bayes' rule gives

$$w(x) = \frac{1-\rho}{\rho}\frac{\eta(x)}{1-\eta(x)}.$$

For a fitted posterior $h(x) \in (0, 1)$, we therefore use

$$\widehat{w}_h(x) = \frac{1-\rho}{\rho}\frac{h(x)}{1-h(x)}.$$

Result 1 (Posterior-to-ratio identity)**.** *We have the exact identity*

$$\mathbb{E}_{P_{\mathrm{b},X}}|w - \widehat{w}_h| = \frac{1}{\rho}\mathbb{E}_M\left[\frac{|\eta - h|}{1-h}\right].$$

[R1]

Let

$$\Delta_{\log}(h) = \mathbb{E}_M\,\mathrm{KL}(\mathrm{Bern}(\eta(X)) \,\|\, \mathrm{Bern}(h(X)))$$

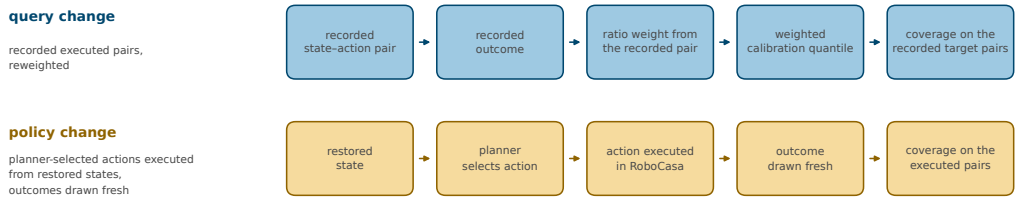


Figure 1: Two kinds of evidence for one statement. Left, query change: the target action law π_t is declared over recorded executed state–action pairs; each recorded outcome was produced once by the physical robot under π_b, and the target law is realized by reweighting the pairs. Right, policy change: a planner enacts π_t from a restored simulator state, and the outcome is drawn fresh under the target law. Both feed the same weighted conformal quantile.

be population binary log-loss regret. The identity converts the ratio error into a pointwise function of (η, h); the regret constrains η through the same pointwise divergence. The tightest computable route from one to the other is the convex dual of that constraint, applied within bins of h so that the regret enters the bound bin by bin, where the posterior sits. For a multiplier $\mu > 0$ and $h \in (0, 1)$ let

$$g_\mu(h) = \sup_{e \in [0,1]} \left\{ \frac{|e - h|}{1 - h} - \mu \, \mathrm{KL}(\mathrm{Bern}(e) \,\|\, \mathrm{Bern}(h)) \right\},$$

the Legendre transform of the divergence at h. Let $B_1, \ldots, B_J$ partition $(0, 1)$ at the M-quantiles of h, and let $\Delta_j = \mathbb{E}_M[\mathbf{1}\{h \in B_j\} \, \mathrm{KL}(\mathrm{Bern}(\eta) \| \, \mathrm{Bern}(h))]$, so that $\sum_j \Delta_j = \Delta_{\log}(h)$.

Result 2 (Per-bin dual bound)**.** *For each fitted posterior $h \in (0, 1)$ and each $J \geq 1$,*

$$\mathbb{E}_{P_{\mathrm{b},X}} |w - \widehat{w}_h| \leq \Pi_J(h) := \frac{1}{\rho} \sum_{j=1}^{J} \inf_{\mu_j > 0} \left\{ \mathbb{E}_M \left[\mathbf{1}\{h \in B_j\} \, g_{\mu_j}(h) \right] + \mu_j \Delta_j \right\},$$

and under the assumptions of P3, with $\widehat{w} = \widehat{w}_h$,

$$\mathbb{P}_{P_b^n \otimes P_t} \{ Y_{n+1} \in C_{\widehat{w}_h}(X_{n+1}) \} \geq 1 - \alpha - \frac{1}{2} \left[\Pi_J(h) + |Z_{\widehat{w}_h} - 1| \right].$$

[**R2**]

Each quantity on the right is computable once h is fitted and Δ_j is available. Where the target law is declared over a finite pool, η is known in closed form and each Δ_j is computed by direct summation. In Section A.6 we state the two looser routes—Cauchy–Schwarz followed by Pinsker, and the dual with a single global constraint—and prove $\Pi_J \leq \Pi_{\mathrm{glob}} \leq \Pi_{\mathrm{CS}}$. Effective sample size and the coverage penalty are controlled by different quantities: the second moment of normalized weights controls efficiency, and the coverage penalty depends on ratio error and the normalizer.

5 Two kinds of evidence

Both experiments evaluate the same statement—P3 applied to the ratio of P2—and differ in how the target law enters the data (Figure 1).

Under a *query change*, the target action law is declared in closed form over recorded executed state–action pairs, and evaluation reweights those pairs. The outcome attached to each pair was produced once, by the physical robot, under the behavior policy; the target law is realized by changing the probability with which each recorded pair is drawn. Physical robot data supplies this kind of evidence and no other, because a recorded transition cannot be re-executed under a different action. The exact ratio follows from the declared law, so learned weighting is compared with oracle weighting, and the coverage bound of R2 is computed with the regret obtained by direct summation.

Under a *policy change*, the target action law is enacted: a planner selects the action at a restored simulator state, the simulator executes it, and the outcome is drawn fresh under the target law. Recorded data cannot supply this evidence, because the target action is not the recorded action. Source and target policies have known densities over a common candidate set, so the propensity ratio is exact, and the comparison is with recalibration on labelled target episodes.

Each kind supplies what the other cannot: the query change tests transport on physical outcomes with a ratio that has to be learned; the policy change tests transport when the outcomes themselves are generated under the target law.

6 Experimental design

Query change. We evaluate Propositions 2 and 3 and R2 with a frozen DINO-feature world model and physical Franka data, using the released JEPA-WMs implementation and DINO-WM checkpoint (FAIR, 2026; Terver et al., 2025; Zhou et al., 2025; Oquab et al., 2023), on the original 192-transition Franka corpus and on FurnitureBench (Heo et al., 2023). The predictor is frozen throughout; we alter only the calibration measure. On the Franka corpus we tilt the whole query law, so the states move with the actions: the tilt is $P_\lambda(i) \propto e^{-\lambda g_i}$ and $Q_\lambda(i) \propto e^{+\lambda g_i}$ over the 192 stored transitions (the source and target query laws P_b and P_t at tilt λ), where $g_i \in [-1, 1]$ is the midrank of the translational action magnitude; g has Pearson correlation 0.80 with the nonconformity score on this pool. Each of the 192 states carries one action, so the state distribution moves exactly as far as the joint law, a total variation of 0.637 at $\lambda = 1.5$. Calibration draws 500 transitions under P_λ and evaluation draws 10,000 under Q_λ, both with replacement from the pool, so 0.33 of a calibration draw is distinct on average, covering 166 of the 192 transitions, and an evaluation draw covers the pool. On FurnitureBench we declare a change in action selection alone, hold the state distribution fixed by construction, and test the construction at the scale of 5,100 episodes with DINOv2. We use one hash-selected one-step query per episode and latent RMSE as the score. We draw the domain-classifier, calibration and evaluation points with replacement from this pool of 5,100 queries and pool coverage over the nine furniture tasks (Appendix C). We tilt the action propensity by $\exp(\kappa b)$ on within-stratum action-magnitude octiles b; the domain classifier receives the stratum indicators and b, the basis in which the true log-ratio is linear. We repeat the Franka experiment with a second frozen encoder, V-JEPA-2-AC (Assran et al., 2025), under the same protocol digest (Appendix D).

Policy change. On RoboCasa (Nasiriany et al., 2024) we use four tasks (CloseFridge, OpenDrawer, TurnOnMicrowave, PickPlaceCounterToCabinet). A latent world model with a frozen DINOv2 encoder and a transformer predictor is trained on 180 demonstration episodes per task and then frozen; a planner enacts each target policy: it scores 32 candidate four-command action chunks under this model, and the chosen chunk is executed from a restored demonstration state, so $d_t = d_b$ and the propensity ratio of P2 is the complete correction. The source policy is ε-randomized demonstration replay ($\varepsilon = 0.5$) with known density over the candidate set; the target policy is a softmax over planner cost with temperature T, with $T = \infty$ the uniform policy and the identical-policy control drawing from the source policy itself. Each task has 150 calibration episodes under the source policy and 150 deployment episodes under each target policy, with one query per episode; we score each chunk the planner chooses against the outcome the simulator returns, by the latent residual of the frozen predictor. Three model seeds are trained; 271828 is primary and the other two are reported as replications. For the recalibration comparison, each task's 150 deployment episodes are divided into 75 adaptation episodes, of which 25, 50, or 75 are labelled and used to recalibrate, and 75 held-out evaluation episodes on which each construction is scored. Intervals are 10,000-draw whole-procedure bootstraps at the episode level. Section C.4 records the configuration.

7 Results

7.1 Query change: learned weights track the oracle on both corpora

On the Franka slice, coverage without weights is 0.902 at $\lambda = 0$ and falls to 0.550 at $\lambda = 1.5$ as the target law separates from the calibration law. Oracle weighting gives 0.910 at the strongest tilt and learned weighting gives 0.909. At FurnitureBench scale, coverage without weights falls to 0.734 at $\kappa = 1.5$; learned weighting gives 0.903 and oracle weighting 0.903. Learned weighting sits -0.0013 from the oracle under the Franka tilt and -0.0008 under the FurnitureBench tilt. Because the FurnitureBench law holds occupancy at one by construction, the occupancy factor of P2 deviates from one by 0 and the product of the two factors reproduces w to within 2.7×10^{-15} at each κ.

With the visual encoder replaced by V-JEPA-2-AC on the Franka slice, holding the windows, tilt laws, seeds, and α grid fixed, coverage without weights is 0.902 at $\lambda = 0$ and 0.561 at $\lambda = 1.5$; learned

Table 1: Each row pairs one question with the experiment that addresses it.

Question	Test	Statistical conclusion
Does source calibration still hold after the query law or the policy changes?	Compare unweighted target coverage across increasing query-change strength and under the planner policy.	Coverage at nominal 0.90 falls to 0.550 (Franka), 0.734 (FurnitureBench), and 0.750 (RoboCasa) while the predictor stays fixed.
Does the identified ratio transport coverage under a query change?	Compare oracle, learned-ratio, and unweighted conformal calibration on both corpora, and on a second visual encoder for the Franka corpus.	Learned weighting returns 0.909 and 0.903, within 0.0013 and 0.0008 of the oracle.
Does exact weighting of the source calibration set replace recalibration under an enacted policy change?	Compare unweighted, exact-weighted, and target-recalibrated coverage on held-out RoboCasa episodes, with an identical-policy control and three model seeds.	Exact weighting reaches 0.912 (0.750 unweighted) with each radius finite; on the held-out subset 0.913 with no new observations against 0.910 with 75 labels per task, and one source set serves all nine target policies where recalibration recollects 300 labelled episodes for each.
What does the bound return under estimation?	Compute R2 on both query-change corpora from the fitted classifier and the declared law.	0.863 and 0.802 against observed 0.909 and 0.903.

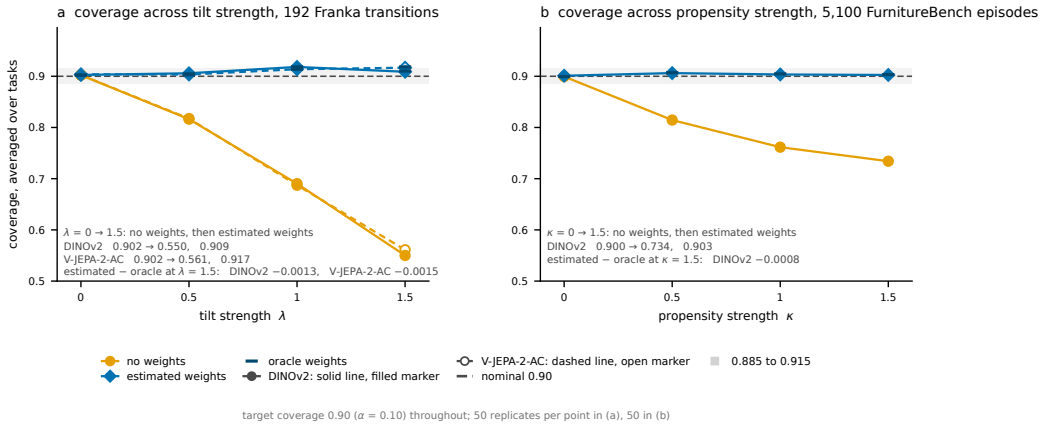


Figure 2: Coverage at target level 0.90 under a query change. (a) The Franka tilt on 192 transitions across tilt strength λ, without weights and with estimated and oracle weights, for DINOv2 and V-JEPA-2-AC. (b) The FurnitureBench propensity tilt on 5,100 episodes across strength κ, averaged over the nine tasks, for DINOv2; the same three constructions.

weighting returns 0.917 and oracle weighting 0.918, against 0.909 and 0.910 on DINOv2. The shift laws are functions of λ and the shift statistic alone, so they are identical across the two encoders; the scores, and with them the coverages, differ.

7.2 Policy change: exact weighting of the source set matches recalibration on target labels

Under the identical-policy control, where the target policy is the source policy and each weight equals one, weighted and unweighted coverage agree exactly, at 0.893, 0.852 and 0.862 on the three model seeds against a finite-sample level of 0.907, and at 0.905, 0.890 and 0.900 on the ten-slot view of the same episodes; no radius is infinite and the effective sample size is 150 calibration episodes per task. We read the weighted results against the seed spread at this control. Under the temperature-3 planner policy, coverage without weights is 0.750; exact propensity weighting gives 0.912 with each radius finite. The paired gain is 16.2 percentage points with a 95% whole-procedure bootstrap interval of $[8.5, 23.7]$, and the mean finite radius rises from 13.95 to 15.95 in the frozen latent metric. On the two further model seeds the same comparison gives 0.690 to 0.877 and 0.735 to 0.892, with 0.852 and 0.862 under the identical-policy control and no infinite radius at temperature 3 on either.

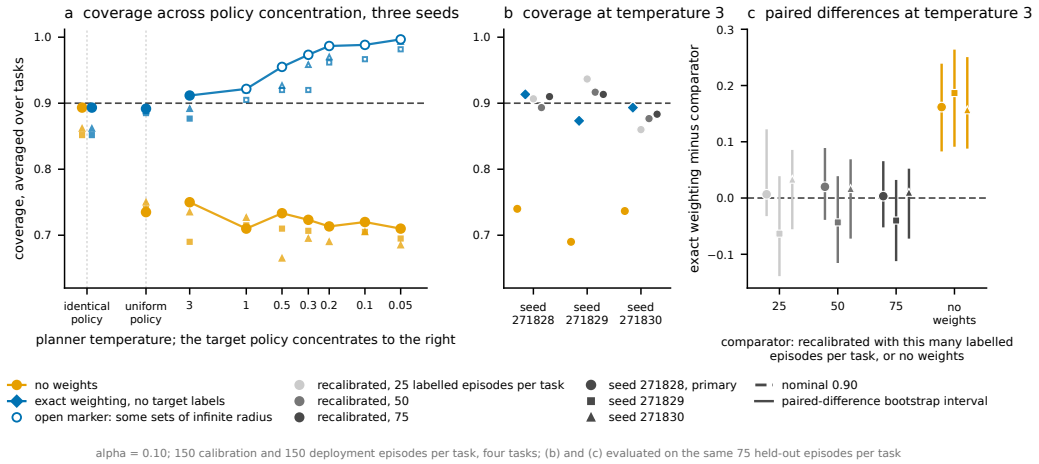


Figure 3: Coverage at target level 0.90 under an enacted policy change on RoboCasa, averaged over four tasks. (a) Coverage across the planner temperature, from the identical-policy control to the uniform policy, for the three model seeds (271828 primary), without weights and with exact propensity weights; open markers mark settings with infinite radii. (b) On the held-out 75-episode subset per task at temperature 3: unweighted, exact weighting with no target labels, and recalibration on 25, 50, and 75 labelled target episodes per task, for the three seeds. (c) Paired differences, exact weighting minus the comparator, with bootstrap intervals, for the primary seed.

The recalibration comparison uses the held-out 75 deployment episodes per task. There, exact weighting of the source calibration set with no target labels gives 0.913 for the primary seed, against 0.740 without weights. Recalibrating on labelled target episodes gives 0.907 at 25, 0.893 at 50, and 0.910 at 75 labels per task. The paired difference between exact weighting and recalibration at 75 labels is +0.3 points with interval $[-5.0, 6.3]$: on this held-out set the source calibration set, exactly reweighted, is a substitute for 75 labelled target episodes per task. The two constructions need different data. Recalibration draws its 300 labelled episodes, 75 per task, under the policy being deployed, so the collection repeats whenever the policy changes: across the nine target policies we evaluate, that is 2,700 labelled deployment episodes per model seed. The weighted construction reuses one 600-episode source calibration set at all nine and requires no observation drawn under any target policy.

The overlap condition. The guarantee of P3 is informative when the target policy places its mass where the source policy has mass. As the planner temperature falls, the softmax concentrates on the planner's preferred chunk and the effective sample size of the weighted calibration set falls: 150 under the identical policy, 70.6 under the uniform policy, 63.6 at temperature 3, 29.9 at temperature 1, 10.0 at temperature 0.5, and 1.8 at temperature 0.05 (Figure 4a). The rate of infinite radii records the same loss of overlap: 0 under the identical policy, the uniform policy, and temperature 3; 0.153 at temperature 1; 0.567 at temperature 0.5; and 0.985 at temperature 0.05. An infinite radius is a measurement that the test point's normalized weight exceeds α, which is the overlap condition violated at that point; the accompanying coverage of 0.997 at temperature 0.05 is coverage by the whole outcome space.

7.3 The coverage bound

We compute R2 with $J = 5$ quantile bins and $\rho = 0.5$ on both query-change corpora, with h the fitted logistic domain classifier of each replicate, Δ_j by direct summation from the declared law, and the normalizer term $|Z_{\widehat{w}_h} - 1|$ from the known pool law, averaged over 50 replicates. Under the Franka tilt at $\lambda = 1.5$ the lower bound is 0.863 against observed coverage 0.909; under the FurnitureBench tilt at $\kappa = 1.5$ it is 0.802 against 0.903 (Figure 4b,c). No score enters the bound, because we build both declared laws from the issued actions and from state strata fixed before the outcome, and the predictor computes neither. The V-JEPA-2-AC run reports coverage on the Franka corpus only: the Franka bound values recomputed against its pool equal the DINOv2 values, and we recomputed its FurnitureBench bound values from the same declared laws with no score input, so the same two values hold on either encoder. The comparison routes of Section A.6 return 0.765 and 0.524 at the same settings.

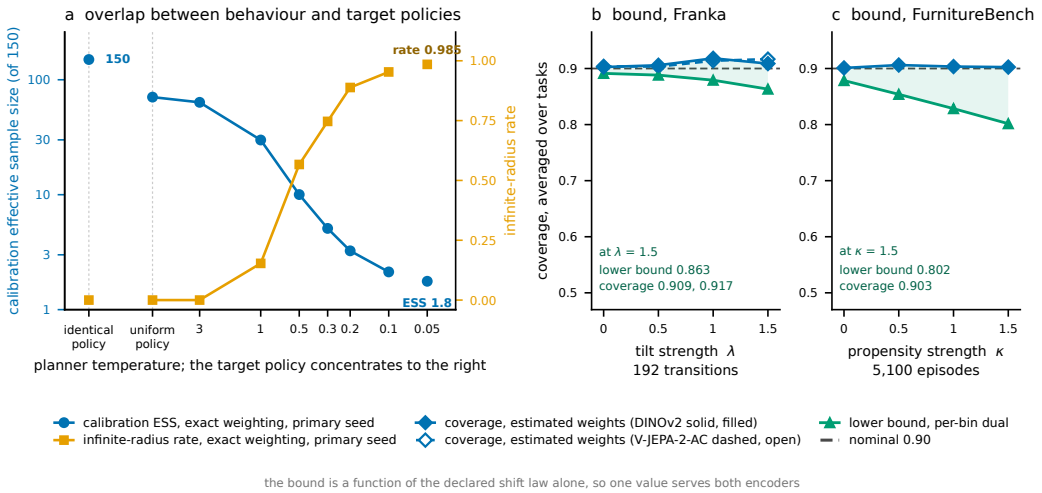


Figure 4: (a) Effective sample size of the weighted calibration set and the rate of infinite radii across the RoboCasa planner temperature, primary seed. (b, c) Observed coverage and the per-bin dual lower bound of R2 across the query-change strength on the Franka tilt (λ) and on the FurnitureBench tilt (κ).

8 Discussion

The world model approximates the conditional transition mechanism; the deployment policy determines which state–action queries receive probability mass. A new policy can therefore change the validity of a calibration statement without changing the predictor. Under a shared conditional mechanism, we identify this change as a likelihood ratio on pre-outcome variables and transport conformal calibration instead of retraining the world model or recalibrating on target labels.

We calibrate the state that the world model predicts after one executed action chunk, and the chunk is a covariate of the query, so we know the weight before we see the outcome. Planners built on learned world models re-plan after each executed action or chunk (Chua et al., 2018; Hansen et al., 2024; Zhou et al., 2025; Assran et al., 2025). When we predict across several re-planned chunks, the later actions depend on states that the prediction describes, so the weight then depends on the outcome (Foffano et al., 2023). Over a rollout the occupancy factor of P2 departs from one, and the variance of the per-step product of propensity ratios grows with the horizon; Liu et al. (2018) and Nachum et al. (2019) avoid that product by estimating occupancy ratios directly. The occupancy factor equals one by design on FurnitureBench and in RoboCasa, where we restore a recorded state at each query; on the Franka corpus it does not, and we declare or learn the ratio as a whole.

The setting is the one in which learned world models are already deployed for action selection: model-predictive control and latent safety filtering over a learned latent state (Chee et al., 2024; Seo et al., 2025), and policy evaluation inside the model itself (Quevedo et al., 2025).

The two kinds of evidence agree. On recorded physical data with a declared target law, learned weighting tracks the oracle at both scales and on both encoders; on simulator data with an enacted target policy, exact weighting of the source calibration set gives the coverage that 75 labelled target episodes per task give by recalibration, and gives it without collecting an observation under the new policy, so the same source set serves each policy the planner is later given. When the policy or occupancy densities are unavailable, we estimate their product from source and target covariates through domain classification, and R2 converts the classifier's regret into a coverage lower bound that sits 0.046 and 0.101 below the observed coverage on the two query-change corpora. The overlap condition is measured rather than assumed: the effective sample size and the rate of infinite radii report, for each target policy, whether the source calibration set has the support that the guarantee needs.

AI use statement

We used a large language model for prose editing, document organization, and grammar assistance, for the conversion of the manuscript to this template, and for the reduction of the main text to the

page limit. A large language model also assisted with the experiment code. The research questions, the experimental designs, the mathematical derivations and the claims are the authors'. We reviewed all AI-assisted text against the results files from which the reported numbers are read, and all AI-assisted code against the outputs it produced. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

Assumptions are stated in Section 2 and complete proofs of P1, P2, P3, and R1 are in Appendix A; Section A.6 states and proves the comparison routes for the ratio-error penalty; Appendix B contains the optimal-transport construction and its proof. Appendix C records the frozen checkpoints, the sampling design of each arm, the hashing rule that selects one query per episode, the FurnitureBench target law, the RoboCasa configuration, and the bootstrap procedures; Appendix D records the second-encoder protocol digest and file hashes. The Lean 4 development that machine-checks the finite-support identities of Appendix E is available at the anonymized commit-pinned repository linked in the abstract footnote and in Appendix E.

References

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016. URL https://arxiv.org/abs/1606.03966.

Christopher Agia. *Deployment-Time Reliability of Learned Robot Policies*. PhD thesis, Stanford University, 2026. URL https://arxiv.org/abs/2603.11400. arXiv:2603.11400.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009. URL https://jmlr.org/papers/v10/bickel09a.html.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 2024. URL https://arxiv.org/abs/2008.04267.

Kong Yao Chee, Thales C. Silva, M. Ani Hsieh, and George J. Pappas. Uncertainty quantification and robustification of model-based controllers using conformal prediction. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 528–540. PMLR, 2024. URL https://proceedings.mlr.press/v242/chee24a.html.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pp.

625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37. URL https://link.springer.com/chapter/10.1007/978-3-030-79876-5_37.

Do Robots Need World Models? Workshop Organizers. Do robots need world models? workshop at the conference on robot learning. https://do-robots-need-world-models.github.io/, 2026. URL https://do-robots-need-world-models.github.io/.

FAIR. JEPA-WMs: Code, data and weights for physical planning with joint-embedding predictive world models, 2026. URL https://github.com/facebookresearch/jepa-wms. Git commit 13cf1d9c7e476f53c17714d2e0f1dc239a883ce0.

Daniele Foffano, Alessio Russo, and Alexandre Proutiere. Conformal off-policy evaluation in markov decision processes. In *IEEE Conference on Decision and Control (CDC)*, 2023. URL https://arxiv.org/abs/2304.02574.

Jason Gauci, Edoardo Conti, Yitao Liang, Kittipat Virochsiri, Yuchen He, Zachary Kaden, Vivek Narayanan, Xiaohui Ye, Zhengxing Chen, and Scott Fujimoto. Horizon: Facebook's open source applied reinforcement learning platform. *arXiv preprint arXiv:1811.00260*, 2018. URL https://arxiv.org/abs/1811.00260.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cf73d57b6dcda32b293df7c2d5341f49-Abstract-Conference.html.

Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.041. URL https://www.roboticsproceedings.org/rss19/p041.html.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026. URL https://arxiv.org/abs/2605.00080.

Yash Jangir, Yidi Zhang, Pang-Chi Lo, Kashu Yamazaki, Chenyu Zhang, Kuan-Hsun Tu, Tsung-Wei Ke, Lei Ke, Yonatan Bisk, and Katerina Fragkiadaki. RobotArena ∞: Scalable robot benchmarking via real-to-sim translation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.23571.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Elis Karcini, Faisal Mehrban, Quang Nguyen, Mac Schwager, Arash Ajoudani, César Cadena, Jan Peters, Marco Hutter, and Haitham Bou-Ammar. Robots need more than VLAs & world models. *arXiv preprint arXiv:2606.06556*, 2026. URL https://arxiv.org/abs/2606.06556.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. URL https://arxiv.org/abs/2403.12945.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://arxiv.org/abs/1604.04173.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*, 83(5):911–938, 2021. doi: 10.1111/rssb.12445. URL https://academic.oup.com/jrsssb/article/83/5/911/7056131.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dda04f9d634145a9c68d5dfe53b21272-Abstract.html.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 304–313. PMLR, 2016. URL https://proceedings.mlr.press/v48/menon16.html.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cf9a242b70f45317ffd281241fa66502-Abstract.html.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024. URL https://arxiv.org/abs/2406.02523.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161. URL https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486. URL https://doi.org/10.1214/14-STS486.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.06177.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025. URL https://arxiv.org/abs/2506.00613.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687. PMLR, 2025. URL https://proceedings.mlr.press/v267/richens25a.html.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2008.07146.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4442–4472. PMLR, 2025. URL https://proceedings.mlr.press/v305/seo25a.html.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4. URL https://doi.org/10.1016/S0378-3758(00)00115-4.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V. Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, 2015. URL https://arxiv.org/abs/1502.02362.

Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31512–31524, 2022. URL https://arxiv.org/abs/2206.04405.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *arXiv preprint arXiv:2512.24497*, 2025. URL https://arxiv.org/abs/2512.24497.

The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pp. 367–381, 2020. doi: 10.1145/3372885.3373824. URL https://arxiv.org/abs/1910.09336.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715. URL https://link.springer.com/book/10.1007/b106715.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. *arXiv preprint arXiv:2102.10439*, 2021. URL https://arxiv.org/abs/2102.10439.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240. PMLR, 2023. URL https://proceedings.mlr.press/v205/wu23c.html.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v206/zhang23c/zhang23c.pdf.

Michelle Zhao, Reid Simmons, Henny Admoni, Aaditya Ramdas, and Andrea Bajcsy. Conformalized interactive imitation learning: Handling expert shift and intermittent feedback. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.08852.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Yanfei Zhou, Lars Lindemann, and Matteo Sesia. Conformalized adaptive forecasting of heterogeneous trajectories. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62002–62056. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24l.html.

A Proofs

A.1 Proof of P1

For each bounded measurable f, we write

$$\begin{aligned}\int f(x,y)\,P_t(\mathrm{d}x,\,\mathrm{d}y) &= \int\left[\int f(x,y)K(\mathrm{d}y\mid x)\right]P_{\mathrm{t},X}(\mathrm{d}x)\\ &= \int \frac{\mathrm{d}P_{\mathrm{t},X}}{\mathrm{d}P_{\mathrm{b},X}}(x)\left[\int f(x,y)K(\mathrm{d}y\mid x)\right]P_{\mathrm{b},X}(\mathrm{d}x)\\ &= \int \frac{\mathrm{d}P_{\mathrm{t},X}}{\mathrm{d}P_{\mathrm{b},X}}(x)f(x,y)\,P_b(\mathrm{d}x,\,\mathrm{d}y).\end{aligned}$$

We therefore identify the Radon–Nikodym derivative with the covariate ratio.

A.2 Proof of P2

We apply P1 with $X=(S,A)$. We factor the state–action marginals as $d_b\pi_b$ and $d_t\pi_t$ and apply the chain rule for Radon–Nikodym derivatives to obtain the occupancy and propensity product.

A.3 Proof of P3

Condition on the independently fitted $\widehat{w}$. Write $\mathcal{D}=\big((X_1,Y_1),\dots,(X_n,Y_n)\big)$ for the calibration sample and $E_{\mathcal{D}}=\{(x,y):y\in C_{\widehat{w}}(x;\mathcal{D})\}$ for the coverage event. Since $\mathrm{d}\widehat{P}/\mathrm{d}P_b=\widehat{w}/Z_{\widehat{w}}$ and the weighted quantile is invariant to global rescaling of the weights, $\widehat{w}$ is the exact ratio for $\widehat{P}$, and Tibshirani et al. (2019) gives $\int \widehat{P}(E_{\mathcal{D}})\,P_b^n(\mathrm{d}\mathcal{D})\ge 1-\alpha$. For each $\mathcal{D}$, $P_t(E_{\mathcal{D}})\ge\widehat{P}(E_{\mathcal{D}})-\mathrm{TV}(P_t,\widehat{P})$; integrating proves the first inequality. For the second,

$$\begin{aligned}2\,\mathrm{TV}(P_t,\widehat{P}) &= \mathbb{E}_{P_b}\left|w-\frac{\widehat{w}}{Z_{\widehat{w}}}\right|\\ &\le \mathbb{E}_{P_b}|w-\widehat{w}|+\mathbb{E}_{P_b}\left|\widehat{w}-\frac{\widehat{w}}{Z_{\widehat{w}}}\right|\\ &= \mathbb{E}_{P_b}|w-\widehat{w}|+|Z_{\widehat{w}}-1|.\end{aligned}$$

Because $\mathbb{E}_{P_b}[w]=1$, the normalizer deviation satisfies $|Z_{\widehat{w}}-1|=|\mathbb{E}_{P_b}[\widehat{w}-w]|\le\mathbb{E}_{P_b}|w-\widehat{w}|$, so $1-\alpha-\mathbb{E}_{P_b}|w-\widehat{w}|$ is a lower bound with no normalizer input; $Z_{\widehat{w}}$ is an expectation under P_b of a known function and is estimable from source covariates alone.

A.4 Proof of R1

Proof. Under the pooling construction, M is the marginal law of X. For each bounded measurable g,

$$\mathbb{E}_M[g(X)\eta(X)]=\mathbb{E}\,[g(X)\mathbf{1}\{D=1\}]=\rho\,\mathbb{E}_{P_{\mathrm{t},X}}[g(X)],$$

so $\eta = \rho\ \mathrm{d}P_{\mathrm{t},X}/\,\mathrm{d}M$ holds M-a.e.; the same computation with $\mathbf{1}\{D = 0\}$ gives

$$1 - \eta = (1 - \rho)\frac{\mathrm{d}P_{\mathrm{b},X}}{\mathrm{d}M} \qquad M\text{-a.e.}$$

If $\eta = 1$ on a set of positive M-measure, that set is $P_{\mathrm{b},X}$-null and $P_{\mathrm{t},X}$-nonnull, which A1 excludes; hence $\eta < 1$ and $\mathrm{d}P_{\mathrm{b},X}/\,\mathrm{d}M > 0$ hold M-a.e. Dividing the two derivatives gives

$$w = \frac{\mathrm{d}P_{\mathrm{t},X}}{\mathrm{d}P_{\mathrm{b},X}} = \frac{\eta/\rho}{(1-\eta)/(1-\rho)} = \frac{1-\rho}{\rho}\frac{\eta}{1-\eta}.$$

Placing w and $\widehat{w}_h$ over the common denominator $(1-\eta)(1-h)$ and using $\eta(1-h) - h(1-\eta) = \eta - h$,

$$w - \widehat{w}_h = \frac{1-\rho}{\rho}\left(\frac{\eta}{1-\eta} - \frac{h}{1-h}\right) = \frac{1-\rho}{\rho}\cdot\frac{\eta - h}{(1-\eta)(1-h)}.$$

We change measure from $P_{\mathrm{b},X}$ to M with $\mathrm{d}P_{\mathrm{b},X}/\,\mathrm{d}M = (1-\eta)/(1-\rho)$:

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{b},X}}|w - \widehat{w}_h| &= \mathbb{E}_M\left[\frac{1-\eta}{1-\rho}\,|w - \widehat{w}_h|\right] \\
&= \mathbb{E}_M\left[\frac{1-\eta}{1-\rho}\cdot\frac{1-\rho}{\rho}\cdot\frac{|\eta - h|}{(1-\eta)(1-h)}\right] \\
&= \frac{1}{\rho}\mathbb{E}_M\left[\frac{|\eta - h|}{1-h}\right],
\end{aligned}$$

where the factor $1-\eta$ cancels because it is strictly positive M-a.e., and $h \in (0,1)$ keeps the remaining integrand measurable and nonnegative. Both sides are finite together. □

A.5 Proof of R2

Proof. If no μ_j makes the bin expectation finite, the infimum is $+\infty$ and the bound holds trivially. By the definition of g_μ as a supremum, $|\eta - h|/(1-h) \le g_\mu(h) + \mu\,\mathrm{KL}(\mathrm{Bern}(\eta)\|\,\mathrm{Bern}(h))$ pointwise for each $\mu > 0$. Multiply by $\mathbf{1}\{h \in B_j\}$, integrate under M, take the infimum over μ_j in each bin, sum over j, and apply R1. The coverage statement combines the total-variation bound of P3 with $\mathbb{E}_{P_{\mathrm{b},X}}|w - \widehat{w}_h| \le \Pi_J(h)$. □

A.6 Comparison routes for the ratio-error penalty

Two further routes convert the same regret into a ratio-error penalty. Both are looser than R2; the first is the form that the Lean development of Appendix E machine-checks.

Result 3 (Localized classifier-regret bound)**.** *For each fitted posterior $h \in (0,1)$,*

$$\mathbb{E}_{P_{\mathrm{b},X}}|w - \widehat{w}_h| \le \Pi_{\mathrm{CS}}(h) := \frac{1}{\rho}\sqrt{\mathbb{E}_M(1-h)^{-2}}\sqrt{\frac{\Delta_{\log}(h)}{2}}.$$

If $h \in [\gamma, 1-\gamma]$, the right-hand side is no larger than

$$\frac{1}{\rho\gamma}\sqrt{\frac{\Delta_{\log}(h)}{2}}.$$

[R3]

Proof. From R1, Cauchy–Schwarz gives

$$\mathbb{E}_{P_{\mathrm{b},X}}|w - \widehat{w}_h| \le \frac{1}{\rho}\sqrt{\mathbb{E}_M(1-h)^{-2}}\sqrt{\mathbb{E}_M(\eta - h)^2}.$$

Binary Pinsker gives $2(\eta - h)^2 \le \mathrm{KL}(\mathrm{Bern}(\eta)\|\,\mathrm{Bern}(h))$ pointwise. Integration proves the first inequality. The clipped bound follows from $(1-h)^{-2} \le \gamma^{-2}$. □

The localized constant is observable under the source–target mixture once we fit h. Because $1 + \frac{\rho}{1-\rho}\widehat{w}_h = (1-h)^{-1}$, we can also write $\mathbb{E}_M(1-h)^{-2} = \mathbb{E}_M(1 + \frac{\rho}{1-\rho}\widehat{w}_h)^2$, which uses the mixture second moment of $(1-h)^{-1}$ in place of the clipping constant γ^{-2}.

Result 4 (End-to-end target coverage)**.** *Under the assumptions of P3, with* $\widehat{w} = \widehat{w}_h$*,*

$$\mathbb{P}_{P_b^n \otimes P_t}\{Y_{n+1} \in C_{\widehat{w}_h}(X_{n+1})\} \geq 1 - \alpha - \frac{1}{2}\left[\frac{1}{\rho}\sqrt{\mathbb{E}_M(1-h)^{-2}}\sqrt{\frac{\Delta_{\log}(h)}{2}} + |Z_{\widehat{w}_h} - 1|\right]. \tag{R4}$$

Result 5 (Global dual bound and ordering)**.** *For each fitted posterior* $h \in (0,1)$*, with* g_μ *as in Section 4,*

$$\mathbb{E}_{P_{b,X}}|w - \widehat{w}_h| \leq \Pi_{\text{glob}}(h) := \frac{1}{\rho}\inf_{\mu>0}\left\{\mathbb{E}_M[g_\mu(h)] + \mu\Delta_{\log}(h)\right\},$$

and $\Pi_J(h) \leq \Pi_{\text{glob}}(h) \leq \Pi_{\text{CS}}(h)$*.* [R5]

Proof. The first inequality is R2 with $J = 1$. For $\Pi_J \leq \Pi_{\text{glob}}$, fix $\mu > 0$ and take $\mu_j = \mu$ in each bin; the per-bin objectives sum to $\mathbb{E}_M[g_\mu(h)] + \mu\Delta_{\log}(h)$ because $\sum_j \Delta_j = \Delta_{\log}(h)$, so the infimum over separate μ_j is no larger than the infimum over a common μ. For $\Pi_{\text{glob}} \leq \Pi_{\text{CS}}$, write $c = (1-h)^{-1}$ and $u = \text{KL}(\text{Bern}(e)\|\,\text{Bern}(h))$. Binary Pinsker gives $|e - h| \leq \sqrt{u/2}$, so $g_\mu(h) \leq \sup_{u\geq 0}\{c\sqrt{u/2} - \mu u\} = c^2/(8\mu)$. Hence $\mathbb{E}_M[g_\mu(h)] + \mu\Delta_{\log}(h) \leq \mathbb{E}_M[c^2]/(8\mu) + \mu\Delta_{\log}(h)$, and the choice $\mu = \sqrt{\mathbb{E}_M[c^2]/(8\Delta_{\log}(h))}$ makes the right-hand side $\sqrt{\mathbb{E}_M[c^2]}\sqrt{\Delta_{\log}(h)/2}$, which is $\rho\,\Pi_{\text{CS}}(h)$. □

On the Franka tilt at $\lambda = 1.5$ the three penalties are $\Pi_J = 0.039$, $\Pi_{\text{glob}} = 0.048$, and $\Pi_{\text{CS}} = 0.089$ with the M-average constant, against an exact L^1 ratio error of 0.038; on the FurnitureBench tilt at $\kappa = 1.5$ they are 0.163, 0.534, and 0.718 against 0.116. Each penalty is at least the exact error on each of the replicates. The coverage levels these comparison routes return are 0.765 on the Franka tilt, where the shipped route used the clipping constant γ^{-2} with $\gamma = 0.05$ in place of the M-average, and 0.524 on the FurnitureBench tilt.

A.7 A FINITE-SAMPLE DOMAIN-RISK INTERFACE

Let $\mathcal{H}$ be a clipped posterior class and let $\widehat{h}$ be a ζ-approximate empirical risk minimizer. If, with probability at least $1 - \delta$,

$$\sup_{h\in\mathcal{H}}|R_{\log}(h) - \widehat{R}_N(h)| \leq u_N(\delta),$$

and $\eta \in \mathcal{H}$, then

$$\Delta_{\log}(\widehat{h}) \leq 2u_N(\delta) + \zeta.$$

Substitution into R5 or R3 yields a high-probability ratio-error bound; the per-bin form R2 takes the vector (Δ_j), which the finite-pool experiments compute from the declared law. For a finite clipped class of size K, Hoeffding's inequality gives

$$u_N(\delta) = \log(1/\gamma)\sqrt{\frac{\log(2K/\delta)}{2N}}.$$

Our finite-pool query-change experiments compute the population regret directly.

B OPTIMAL TRANSPORT DOES NOT IDENTIFY THE RATIO

We construct distributions for which a small optimal-transport distance does not control the likelihood ratio required by weighted conformal prediction. Fix $a \in (0,1)$ and let P_n be uniform on $[0,1]$. Partition the interval into $2n$ adjacent cells of width $1/(2n)$ and let Q_n have density $1 + a$ on even cells and $1 - a$ on odd cells.

Proposition 4 (Small transport distance, non-vanishing ratio error)**.** *We construct the sequence so that*

$$\left\|\log\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right\|_\infty \leq \max\{\log(1+a), -\log(1-a)\},$$

$$\chi^2(Q_n\|P_n) = a^2, \qquad \text{TV}(Q_n, P_n) = \frac{a}{2},$$

and

$$W_1(P_n, Q_n) \le \frac{a}{4n}, \qquad W_2(P_n, Q_n)^2 \le \frac{a}{8n^2}. \tag{P4}$$

Proof. We use the likelihood ratio $1 \pm a$ and direct integration to obtain the log-ratio, χ^2, and total-variation identities. We couple the common mass at each point and move the excess a-fraction from each odd cell to the adjacent even cell. We move total mass $a/2$ by distance $1/(2n)$, which gives the two transport bounds. □

We therefore obtain $W_1, W_2 \to 0$ while total variation and χ^2 remain fixed. Optimal transport controls Lipschitz expectations under additional regularity (Villani, 2009); it does not identify the covariate-shift ratio.

C Experimental details

Figure 5 gives the call structure, in which ratio-train and ratio-audit fit and audit h, calibration supplies the scores entering the weighted quantile, and test supplies the evaluation points. The RoboCasa policy change realises this partition at the episode level: world-model training, calibration and deployment episodes are disjoint, 180, 150 and 150 per task. The two query-change experiments are resampling studies on a fixed scored pool. During deployment we observe each outcome one chunk after its query, so we can also track deployed coverage on the stream and test sequentially for a change (Podkopaev & Ramdas, 2022; Vovk et al., 2021).

C.1 Frozen model and scores on the query-change corpora

We use the released DINO-WM checkpoint trained on DROID and RoboCasa data. We evaluate it on FurnitureBench and the released Franka evaluation slice. For each transition we encode two context frames, apply the issued seven-dimensional action sequence, unroll one step, and compare the predicted latent with the observed next-frame DINOv2 latent. We use latent RMSE as the score. On the Franka slice at $\alpha = 0.10$, the mean radius under learned weighting rises from 1.18 at $\lambda = 0$ to 1.24, 1.27 and 1.28 at $\lambda = 0.5$, 1 and 1.5, within 0.001 of the oracle radius at each tilt, while the unweighted radius falls to 0.94 as its coverage falls. The Franka domain classifier is a logistic regression in g alone, fitted per replicate by L-BFGS-B on 4,000 draws under P_λ and 4,000 under Q_λ from the 192 transitions with a ridge penalty of $\frac{1}{2}\,10^{-6}$ on the slope; the posterior is clipped to $[0.05, 0.95]$ before forming $\widehat{w} = h/(1-h)$ at $\rho = \frac{1}{2}$, the clipped variant caps $\widehat{w}$ at 5, and the regret is computed exactly over the pool against $\eta = Q_\lambda/(P_\lambda + Q_\lambda)$.

C.2 FurnitureBench sampling

FurnitureBench contains nine furniture tasks. We use all 5,100 episodes. We score eight evenly spaced windows per episode at 5 Hz and select one primary window per episode by hashing the episode identifier. The declared-propensity arm draws from this fixed pool of 5,100 primary windows: per replicate it fits the domain classifier on 4,000 draws under the source law and 4,000 under the target law, then draws 500 calibration points under the source law and 10,000 evaluation points under the target law, with replacement, and pools coverage over the evaluation draws. We report the mean and Monte-Carlo standard error over 50 replicates. Ratio fitting and calibration therefore draw from one pool rather than from disjoint episode blocks.

C.3 Query change on FurnitureBench

We discretize state into twelve strata and action magnitude into eight within-stratum bins. We construct the behavior law to reproduce the uniform primary-episode pool. We hold state occupancy fixed and tilt only the action-bin propensity by $\exp(\kappa b)$ for $\kappa \in \{0, 0.5, 1, 1.5\}$. The resulting log density ratio is linear in action-bin index and stratum indicators, so the domain classifier is correctly specified. Coverage is averaged over 50 replicates per κ.

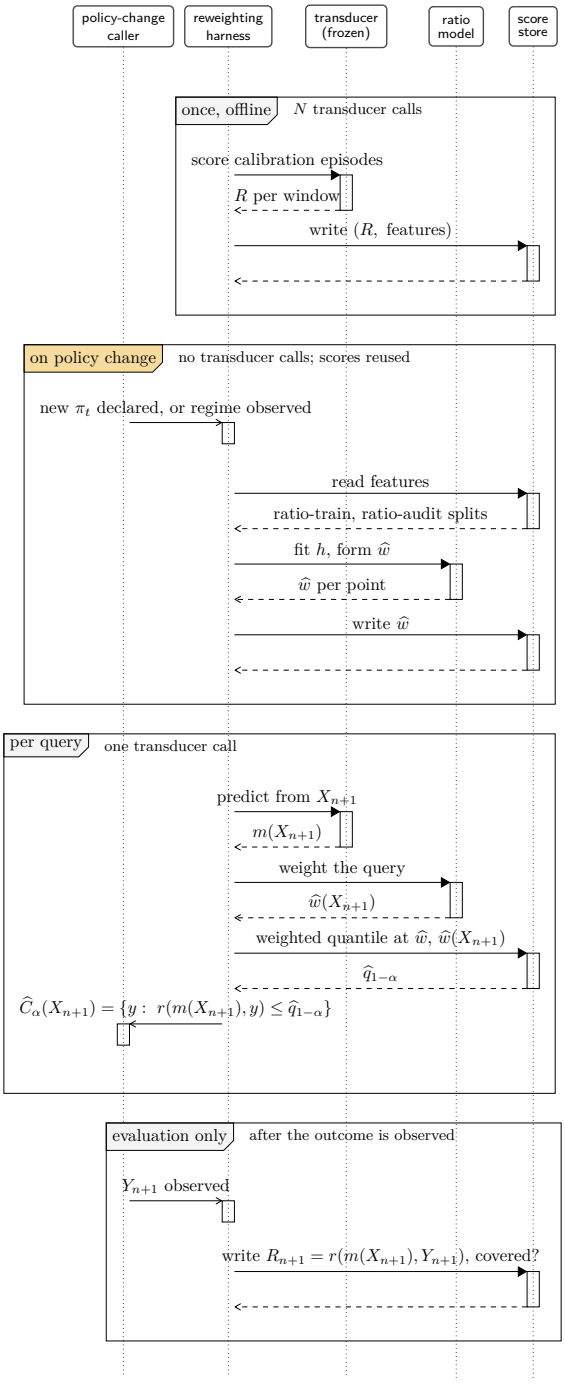


Figure 5: Call structure of the reweighting harness. Scores and ratio features are written once per calibration episode. A policy change reaches the harness as a declared or enacted π_t; it reweights the stored calibration set and triggers a refit of h on the ratio-train and ratio-audit splits, or a direct evaluation of the known propensity ratio, and a rewrite of $\widehat{w}$, with no call to the transducer. We answer a query with one transducer call returning the prediction $m(X_{n+1})$, the ratio model's weight $\widehat{w}(X_{n+1})$ for the test point, and a weighted-quantile lookup against the stored scores that uses both; the outcome Y_{n+1} enters only after the set is returned, when its score is written for evaluation.

C.4 Policy change on RoboCasa

Tasks: CloseFridge, OpenDrawer, TurnOnMicrowave, PickPlaceCounterToCabinet (RoboCasa 1.0.1, robosuite 1.5.2, MuJoCo 3.3.1). Encoder: `facebook/dinov2-small`, frozen. Predictor: a

transformer of width 384, six layers, six heads, MLP ratio 4, dropout 0.1, trained for 200,000 updates at batch size 64, learning rate 3×10^{-4}, weight decay 0.01, on 180 training and 20 validation demonstration episodes per task, for each of the model seeds 271828, 271829, and 271830. Control at 20 Hz with context stride 2 from the `robot0_agentview_left` camera at 224 pixels. Each query restores a demonstration state (state tolerance 10^{-8}, pixel tolerance 1), the planner scores 32 candidate chunks of four commands, and the selected chunk is executed. The candidates are the logged chunk and 31 scrambled-Sobol Gaussian perturbations of it in normalized controller-input units, with standard deviation 0.25 of the input half-range on the ten continuous channels, the gripper command flipped with probability 0.25, and all values clipped to the controller bounds. The planner cost of a candidate is the mean squared distance between the latent the frozen predictor returns after the chunk and the encoded final frame of the demonstration; costs are standardized across the 32 candidates and the target policy is the softmax of the negative standardized cost at temperature T. At $T = 3$ and $\alpha = 0.10$ on the held-out 75 episodes per task the mean finite radius is 15.96 for the exact-weighted source set and 16.18, 15.87 and 16.00 after recalibration on 25, 50 and 75 labels; on the two further model seeds it is 15.76 against 16.27 and 15.73 against 15.81 at 75 labels, so the reweighted sets are no wider than the recalibrated ones. At 75 labels the recalibrated comparator reaches 0.913 and 0.883 on the two further seeds against exact weighting at 0.873 and 0.893, with paired differences of -4.0 and $+1.0$ points and intervals $[-11.0, 3.0]$ and $[-7.0, 5.0]$. Source policy: ε-randomized demonstration replay with $\varepsilon = 0.5$. Target policies: softmax over planner cost at temperatures ∞, 3, 1, 0.5, 0.3, 0.2, 0.1, 0.05, and the identical-policy control. Episodes per task: 150 calibration, 150 deployment; ten query slots per episode with slot 0 primary. Nominal levels 0.10, 0.05, 0.20, with 0.10 reported. Bootstrap: 10,000 whole-procedure repetitions (seed 141421). Recalibration comparison: 75 adaptation and 75 evaluation episodes per task, adaptation sizes 25, 50, 75. Split seed 271828, candidate seed 314159, sampling seed 161803. Protocol SHA-256 `27a23d12fdad564def2a7001b9e7e30d9545c80a90b8c072acaa24703cddb439`. A secondary construction with a fixed weight bound of 64 returns an infinite radius at each of the 81 settings we ran, across three nominal levels, nine target policies and three model seeds, and is not reported further.

D Second visual encoder and file digests

We repeat the Franka query-change experiment with the frozen DINOv2 transducer replaced by V-JEPA-2-AC, and hold everything else fixed: the same protocol digest `4ff1c7015251dd09 6496710b55b63da23128aff35a6074a4eb401c278552358f`, the same windows, tilt laws, seeds and α grid. The world model is `droid_vj2ac_noprop.pth.tar` (SHA-256 `c08425152bf3eee07641654511666c63b1117344328908839 89ad2d9b3ba3a6`) with visual encoder `vjepa2_vit_giant.pth` (SHA-256 `67129f011434e605 d894e69f2c8e13d9db118deabe59d54bf6e0fa62c2c5cb8e`), at implementation commit `13cf1d9c`. Because the declared laws are functions of the shift strength and the shift statistic, the shift-law table is bit-identical across the two encoders, and on the FurnitureBench law the occupancy and factorization identities hold at deviation 0 and at most 2.7×10^{-15} at each κ; on that corpus the V-JEPA-2-AC run produced the shift-law checks, and its coverage is reported for the Franka slice. Each encoder scores frames decoded at its native training resolution, 224 by 224 for DINOv2 and 256 by 256 for V-JEPA-2-AC, under the same window rule and frame indices. V-JEPA-2-AC layer-normalizes prediction and target, so score scales and hence absolute radii are not comparable across encoders; coverage is the cross-encoder comparand. No score enters the query-change bound, because we build both declared laws from the issued actions and from state strata fixed before the outcome, and the predictor computes neither. The V-JEPA-2-AC run reports coverage on the Franka corpus only: the Franka bound values recomputed against its pool equal the DINOv2 values, and we recomputed its FurnitureBench bound values from the same declared laws with no score input.

E Machine-checked identities

We formalize finite-support versions of Results 1 and 3 and P2 and the arithmetic composition with P3 in Lean 4 and Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). We machine-check the posterior-odds cancellation, binary Pinsker step, weighted Cauchy–Schwarz step,

occupancy–propensity identity, balanced-domain specialization, and coverage-bound composition. We take weighted conformal validity as the cited premise from Tibshirani et al. (2019) and do not re-formalize it. The development comprises two self-contained Lean packages, one statistical and one causal, available at the anonymized commit-pinned repository, which maps each result to its declaration. Each builds against its own commit-pinned `mathlib4` revision and requires no further dependency; the two pins are incompatible, so the packages are built separately rather than merged.