EPISTEMIC ACTION QUOTIENTS: A THEORY OF CONSEQUENCE-BASED ACTION ABSTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

When should physically different sensing actions be treated as the same? We argue that the answer depends on what the robot needs to learn rather than on how the actions move it. We introduce Epistemic Action Quotients (EAQs), which define action identity through task-conditioned epistemic consequences: two actions are exactly equivalent for a demand when they induce the same distribution over task-projected belief updates. For finite data, we use bounded-diameter operational partitions and establish a Lipschitz within-block regret bound together with a sufficient stability condition under bounded drift. In the tested HM3D and Habitat settings, epistemic identity repeatedly differs from motor similarity and changes with the demand; a sensor pan and a whole-body yaw fall into the same class in all 20 seeds. In a controlled planning study, evaluating one representative per class reduces the candidate set eightfold with no observed regret. We further instantiate Online EAQ, which predicts action consequences from locally available pre-action information and treats demand changes as mandatory refresh events. In an HM3D pilot, task-triggered refresh accounts for most of the improvement over static caching, while all 25 transitions satisfying the realized-branch analogue of $2\eta_t < \rho_t$ preserve the reference partition.

1 INTRODUCTION

A mobile robot that must check whether the corridor ahead is clear can pan its camera, yaw its whole body, or take a small step forward. These actions use different actuators and differ in energy, time, and collision risk, yet for this question they may yield the same evidence (Figure 1a). If the robot instead asks whether the space to its left is free, its actions group differently. Which sensing actions are interchangeable therefore depends on what the robot is trying to learn, not on how the actions move it.

Identifying interchangeable actions matters because selecting a sensing action can be expensive. Active perception and value-of-information methods score each candidate by predicting what it would reveal (Ghasemi et al., 2020), so their cost grows with the number of candidates. If interchangeable actions are known in advance, a planner needs to score only one representative per class. Whether this reduction is safe, however, depends on the grouping criterion. Grouping by motor similarity can merge actions that reveal different information and separate actions that reveal the same, so a planner that relies on it may discard the most informative action.

Existing notions of action equivalence do not capture this dependence on the question. MDP homomorphisms and bisimulation identify states or actions that share reward and transition structure (Ravindran, 2004; Givan et al., 2003; Ferns et al., 2011), and action-grouping and state–action equivalence methods rely on the same reward–transition behavior (Asadi et al., 2019; Li et al., 2024). These criteria compare actions by how they change the world, whereas a sensing action is valuable for what it reveals about it. Learned action representations place actions close together when their behavior or motor effects are similar (Chandak et al., 2019; Gu et al., 2022), and language- or task-conditioned latent actions (Karamcheti et al., 2022; Ye et al., 2025) condition the representation on the task without specifying when two actions are interchangeable. For sensing actions, the relevant criterion is instead the information an action provides about the robot's current question.

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.

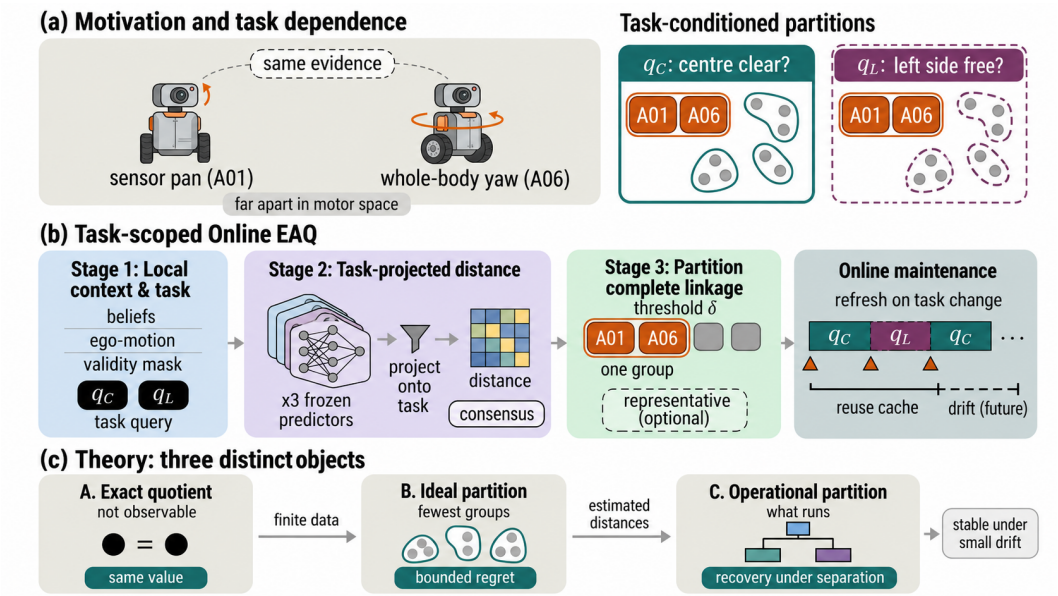


Figure 1: **Overview of EAQ and the evaluated Online EAQ operating point.** (a) A sensor pan (A01) and a whole-body yaw (A06) are far apart in motor space but fall into the same class under both demands q_C and q_L, while the other actions regroup with the demand (schematic). (b) The evaluated Online EAQ policy treats demand changes as mandatory refresh events and otherwise reuses the cached partition; fixed-demand reuse is justified only when the stability condition holds. (c) Three levels of action identity and the guarantee attached to each.

We propose Epistemic Action Quotients (EAQs), which define action identity by task-conditioned epistemic consequences. A demand q specifies which part of the robot's belief is relevant to its current question, and two actions are exactly equivalent for q if they induce the same distribution over belief updates projected onto that part. Because exact equality of distributions cannot be verified from finite data, we introduce an ideal minimum-cardinality δ-partition and the deterministic complete-linkage partition used in practice, and establish which guarantees hold for each. The definition specifies *what* makes two actions interchangeable at a single decision, but not *when* a cached operational partition remains valid. A demand change replaces the projection M_q and therefore requires the quotient to be reconsidered. When the demand remains fixed, reuse is only conditionally justified: our stability theorem gives a sufficient regime in which drift is small relative to the separation margin. We use *Online EAQ* for the task-scoped dynamic realization evaluated later; its deliberately simple operating point refreshes at demand changes, while the stability diagnostic studies when additional fixed-demand reuse is supported.

Our contributions are as follows:

1. We define a task-conditioned notion of action identity for sensing actions, together with its ideal and operational approximations, and establish value-equivalence, regret, and stability guarantees for these relations.

2. We show in HM3D and Habitat that task-conditioned epistemic equivalence occurs repeatedly in the tested embodied settings, changes with the demand, and groups actions that motor similarity separates.

3. We instantiate Online EAQ as a stand-alone learned consequence-prediction module. In HM3D, task-triggered refresh accounts for most of its improvement over static caching, while a theory-aligned diagnostic empirically supports a conservative fixed-demand reuse region suggested by our drift theorem.

2 Related Work

Action abstraction and equivalence. MDP homomorphisms map actions across states while preserving reward and aggregated transition structure (Ravindran, 2004), and bisimulation identifies states with equivalent reward and transition behavior (Givan et al., 2003; Ferns et al., 2011). Bisimulation metrics have also been used to learn task-relevant state representations from pixels (Zhang et al., 2021). Approximate abstractions relax exact equivalence and bound the resulting loss in value (Abel et al., 2016; 2020), and action-grouping and state–action equivalence methods exploit shared reward–transition behavior to improve learning efficiency (Asadi et al., 2019; Li et al., 2024). EAQ shares the goal of value-preserving compression, and its regret bound is analogous to the value-loss bounds of approximate abstraction. It differs in what is compared: EAQ compares sensing actions by their epistemic consequences rather than by reward and transitions, and conditions this comparison on the current demand.

Action representations. Learned action embeddings place actions that appear in similar contexts or produce similar effects near one another (Tennenholtz & Mannor, 2019; Chandak et al., 2019), and action pseudometrics support offline reinforcement learning over large action sets (Gu et al., 2022). LILA and LAPA further condition latent actions on language or task information (Karamcheti et al., 2022; Ye et al., 2025). These methods learn a geometry over actions from behavioral data. EAQ instead defines an explicit equivalence relation whose classes change with the demand and come with a value guarantee.

Active perception and value of information. Active perception selects sensing actions by their expected contribution to a task (Bajcsy, 1988; Aloimonos et al., 1988), for example through information-gain criteria for next-best-view planning (Isler et al., 2016) or task-oriented perception planning under partially known semantics (Ghasemi et al., 2020). In partially observable settings, these choices are made in belief space (Kaelbling et al., 1998). Value-of-information theory quantifies what an observation is worth for a decision (Howard, 1966), and Bayesian experimental design measures the information an experiment is expected to provide (Lindley, 1956; Chaloner & Verdinelli, 1995). These approaches rank or select actions, whereas EAQ groups candidate actions whose consequences are interchangeable for the current demand, so that a selector needs to evaluate only one representative per class. The closest classical notion is Blackwell's comparison of experiments (Blackwell, 1951), under which two experiments are equivalent if they are equally valuable for every decision problem. EAQ requires equivalence only with respect to the current demand, which yields coarser classes that change as the demand changes.

3 Task-Conditioned Action Identity

Let $\mathcal{A}$ be a finite set of candidate sensing actions. Before acting, the agent holds pre-action information e, and a demand q specifies a task projection M_q that retains the part of the belief relevant to the current question. For each action a, let Y_a^q denote the resulting task-projected belief update, which may be stochastic, and let $P_a^M := P_{a|e}^q = \mathrm{Law}(Y_a^q \mid e)$, with $M = M_q$, denote its distribution on a consequence space $(\mathcal{E}_M, d_M)$. We suppress fixed e, write $\Pi(\mathcal{A})$ for partitions of $\mathcal{A}$, and call $U : \mathcal{E}_M \to \mathbb{R}$ a *consequence-only utility*.

3.1 Exact Equivalence

Definition 3.1 (Exact epistemic equivalence)**.** For $a, a' \in \mathcal{A}$,

$$a \equiv_{e,M} a' \quad \Longleftrightarrow \quad P_a^M = P_{a'}^M. \tag{1}$$

The Exact Epistemic Quotient is $\mathcal{Q}_{e,M}^\star := \mathcal{A}/\equiv_{e,M}$.

Proposition 3.2 (Exact epistemic value equivalence)**.** *If $a \equiv_{e,M} a'$, then every consequence-only utility U satisfies*

$$\mathbb{E}_{P_a^M}[U] = \mathbb{E}_{P_{a'}^M}[U]. \tag{2}$$

Thus any consequence-only objective can evaluate one member per exact class without loss.

3.2 Approximate Equivalence

For finite data we compare consequence laws by Wasserstein-1 distance (Villani, 2009):

$$D_M(a, a') = W_{1,d_M}(P_a^M, P_{a'}^M), \qquad \operatorname{diam}_M(\tau) = \max_{a,a' \in \tau} D_M(a, a'). \tag{3}$$

A partition is *population-δ-feasible* when every block has diameter at most δ under the true laws.

Definition 3.3 (Ideal minimum-cardinality δ-partition)**.** An ideal δ-partition is any solution of

$$\mathcal{P}_M^{\star,\delta} \in \arg\min_{\mathcal{P} \in \Pi(\mathcal{A})} |\mathcal{P}| \quad \text{subject to} \quad \max_{\tau \in \mathcal{P}} \operatorname{diam}_M(\tau) \le \delta. \tag{4}$$

Because the ideal partition need not be unique and exhaustive search is intractable, we use deterministic complete linkage.

Definition 3.4 (Deterministic complete-linkage partition)**.** Given a symmetric dissimilarity matrix G over $\mathcal{A}$, a threshold δ, and a numerical tolerance $\epsilon_{\text{alg}} \ge 0$, complete linkage starts from singleton blocks and repeatedly merges the pair of blocks with the smallest linkage distance

$$d_{\text{CL}}^G(C, C') = \max_{a \in C,\, a' \in C'} G(a, a') \tag{5}$$

as long as this distance is at most $\delta + \epsilon_{\text{alg}}$, breaking ties by a fixed action order. We denote its output by $\text{CL}_{\delta,\epsilon_{\text{alg}}}(G)$.

Complete linkage need not attain the minimum cardinality, but recovers any target partition separated in its input (Appendix A.1):

$$\begin{aligned} \max_{a,a' \in \tau} G(a, a') &\le \delta + \epsilon_{\text{alg}} && \text{within each target block,} \\ \min_{a \in \tau_i, a' \in \tau_j} G(a, a') &> \delta + \epsilon_{\text{alg}} && \text{across distinct target blocks.} \end{aligned} \tag{6}$$

3.3 Regret and Stability Guarantees

Theorem 3.5 (Within-block regret bound)**.** *Let U be L-Lipschitz with respect to d_M, and let $V_M(a) = \mathbb{E}_{P_a^M}[U]$. For every block τ of a partition $\mathcal{P}$, every $\tilde{a} \in \tau$, and $a_\tau^\star \in \arg\max_{a \in \tau} V_M(a)$,*

$$V_M(a_\tau^\star) - V_M(\tilde{a}) \le L \operatorname{diam}_M(\tau). \tag{7}$$

In particular, every block of a population-δ-feasible partition has consequence-only regret at most $L\delta$.

Hence δ trades candidate reduction against worst-case regret; estimated distances enlarge the bound by their estimation error (Corollary B.1; Proposition D.2).

Let D_t be the distance matrix at decision t. Embed consecutive consequence spaces isometrically in $(\bar{\mathcal{E}}, \bar{d})$ and write $Q_{a,t}$ for action a's embedded law.

Assumption 3.6 (Separation margin)**.** A target partition $\mathcal{T}_t$ has separation margin $\rho_t > 0$:

$$\max_{a,a' \in \tau} D_t(a, a') \le \delta - \rho_t \qquad \text{within target blocks,} \tag{8}$$

$$\min_{a \in \tau_i, a' \in \tau_j} D_t(a, a') \ge \delta + \rho_t \qquad \text{across target blocks.} \tag{9}$$

Theorem 3.7 (Partition stability under drift)**.** *Under Assumption 3.6, suppose every action's distribution drifts by at most η_t between consecutive decisions,*

$$W_{1,\bar{d}}(Q_{a,t}, Q_{a,t+1}) \le \eta_t, \tag{10}$$

with $2\eta_t < \rho_t$. Then the ideal and population complete-linkage partitions at t and $t+1$ all coincide with $\mathcal{T}_t$:

$$\mathcal{P}_t^{\star,\delta} = \mathcal{P}_{t+1}^{\star,\delta} = \mathcal{P}_{t,\text{CL}}^{\delta} = \mathcal{P}_{t+1,\text{CL}}^{\delta} = \mathcal{T}_t. \tag{11}$$

The theorem gives a sufficient safe-reuse condition under a fixed demand. Demand changes instead replace M_q and trigger mandatory reconsideration. The evaluated Online EAQ uses demand change as its observable refresh trigger; Section 5.3 evaluates the theorem-aligned reuse condition separately. Predicted-distance extensions and proofs are in the appendix.

Algorithm 1 Task-scoped Online EAQ. The Static baseline skips the refresh step and keeps its initial partition for each demand throughout the episode.

Require: predictors $f^{(1)}, f^{(2)}, f^{(3)}$; demand set $\mathcal{Q}$; threshold δ; tolerance ϵ_{alg}

observe e_0
for each demand $q \in \mathcal{Q}$ **do** ▷ initialization
    $\mathcal{C}[q] \leftarrow \mathrm{CL}_{\delta,\epsilon_{\text{alg}}}\big(\widehat{d}^{(1)}_{0,q}\big)$
end for
for each decision $t = 0, 1, 2, \ldots$ **do**
    observe e_t and q_t
    if $t > 0$ **and** $q_t \neq q_{t-1}$ **then** ▷ refresh step
        compute $\widetilde{D}_t$ from $\widehat{d}^{(1)}_{t,q_t}, \widehat{d}^{(2)}_{t,q_t}, \widehat{d}^{(3)}_{t,q_t}$ and $\mathcal{C}[q_t]$ by Equation (13)
        $\mathcal{C}[q_t] \leftarrow \mathrm{CL}_{\delta,\epsilon_{\text{alg}}}\big(\widetilde{D}_t\big)$
    end if
    output $\widehat{\mathcal{P}}_t \leftarrow \mathcal{C}[q_t]$
end for

4 Task-Scoped Online EAQ

We instantiate EAQ dynamically as Online EAQ, which maintains sensing-action partitions from pre-action information e_t and demand q_t without simulator state or reference partitions. It combines consequence prediction, predictor consensus, and task-triggered refresh (Algorithm 1).

Estimating the distance D_M of Equation (3) directly would require executing every candidate action, so Online EAQ instead predicts each action's consequence from pre-action information. At decision t, the predictor takes features of e_t, namely the current belief over depth bins in three image sectors (left, center, right), its change since the previous decision, the observed ego-motion, and a validity mask, together with a one-hot encoding of the action, and outputs a predicted post-action belief for each sector. For demand q, let $b_{t,q}$ be the current projected belief and $\widehat{b}_{t,a,q}$ the predicted projected belief after action a. We compare actions through their predicted updates $\widehat{c}_{t,a,q} = \widehat{b}_{t,a,q} - b_{t,q}$:

$$\widehat{d}_{t,q}(a, a') = \tfrac{1}{2}\|\widehat{c}_{t,a,q} - \widehat{c}_{t,a',q}\|_1. \tag{12}$$

Because the current belief cancels, this is the total-variation distance between predicted projected beliefs and serves as a surrogate for D_M; Proposition D.2 relates the two. The two-layer MLP is trained offline with belief-prediction and pairwise-relation losses on disjoint episodes and frozen at runtime (Section L).

To reduce seed-specific partition changes, three independently seeded predictors are aggregated according to the cached relation:

$$\widetilde{D}_t(a, a') = \begin{cases} \max_j d^{(j)}_t(a, a'), & \text{if } a, a' \text{ are currently separated,} \\ \mathrm{median}_j\, d^{(j)}_t(a, a'), & \text{if } a, a' \text{ are currently merged.} \end{cases} \tag{13}$$

With $\theta = \delta + \epsilon_{\text{alg}}$, merging a separated pair requires all predictors to agree, whereas one dissenting predictor cannot split a cached pair. This asymmetric rule provides hysteresis; complete linkage is then applied with $\delta = 0.05$.

It remains to decide when to recompute the partition. Action identity depends on both e_t and q_t, but the two affect it in different ways. A new demand replaces the projection M_q and hence the task-projected consequence space and induced action geometry, so a partition computed for one demand carries no guarantee for another. A change in e_t under a fixed demand only moves the consequence distributions within the same space, and Theorem 3.7 shows that the partition remains unchanged as long as per-action drift stays below half the separation margin. The evaluated Online EAQ operating point therefore keeps one cached partition per demand, treats a demand switch as a mandatory refresh event, and otherwise reuses the cached partition; between switches, it performs no prediction at all. This task-triggered operating point is deliberately simpler than a deployable stability-certificate policy: it does not estimate η_t or ρ_t online. Section 5.3 separately tests whether realized-branch analogues of the sufficient stability condition identify fixed-demand transitions on which the reference

Table 1: Controlled Habitat study: number of seeds (out of 20) in which each event occurs, with two-sided Wilson 95% intervals. The task-agnostic partition averages the distances over all three sectors.

Event	q_C	q_L
A01 and A06 in the same class	20/20 [0.839, 1]	20/20 [0.839, 1]
Task-conditioned $\neq$ motor partition	20/20 [0.839, 1]	20/20 [0.839, 1]
q_C partition $\neq$ q_L partition	20/20 [0.839, 1]	
Task-conditioned $\neq$ task-agnostic partition	19/20 [0.764, 0.991]	8/20 [0.219, 0.613]

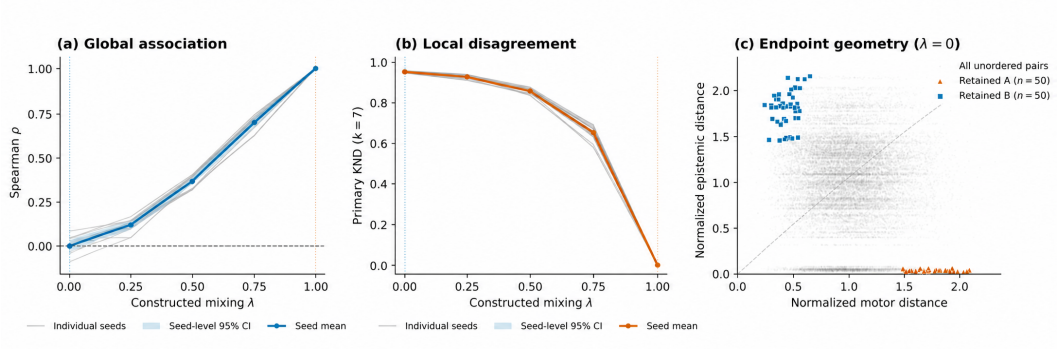


Figure 2: **Motor similarity and epistemic identity can diverge.** (a, b) As motor descriptors are interpolated from class-independent ($\lambda = 0$) to class-aligned ($\lambda = 1$), the rank correlation between motor and epistemic pairwise distances rises and k-nearest-neighbor disagreement falls. (c) At $\lambda = 0$, highlighted pairs have near-zero epistemic distance but large motor distance (A), or large epistemic distance but small motor distance (B). Details are in Section G.

partition is unchanged. Because the consensus rule requires an existing partition, the initial partitions are computed from a single predictor at the first decision of each episode. Section 5.1 shows that a demand switch is more likely to change the reference partition than a step under the same demand, and Section 5.3 examines whether refreshing at demand switches improves agreement with it.

5 Experiments

We evaluate three questions. (Q1) Does task-conditioned identity arise in embodied sensing, and does it differ from motor similarity? (Q2) When the ground truth is known, does complete linkage recover it, and does grouping by consequence preserve decision quality? (Q3) Does refreshing at demand changes improve the partitions maintained online, which component produces the gain, and does the stability condition identify safe fixed-demand reuse? Sections 5.1 to 5.3 address them in turn.

5.1 Task-Conditioned Identity in Embodied Sensing

Reference partitions in HM3D. We first measure action identity without any learned model. Over 12 HM3D episodes (360 decisions), we restore the simulator state at each decision, execute each of the seven actions once, and partition the actions by their task-projected belief updates. The reference partition has fewer classes than actions at every decision. The partitions for q_C and q_L differ at 270 of 360 decisions (75%), whereas under a fixed demand the partition changes at 132 of 324 consecutive transitions (41%). A demand switch is thus nearly twice as likely to change action identity as a step under the same demand, although within-demand changes remain frequent.

A controlled pan–yaw comparison. To isolate the example of Figure 1, we use a single Habitat scene with the same seven actions and 20 seeds, each sampling a different pose (Section J). A 10° sensor pan (A01) and a 10° whole-body yaw (A06) use different actuators and are separated by the motor partition, yet at $\delta = 0.05$ they fall into the same class under both demands in all 20 seeds

Table 2: Synthetic studies with known consequence classes (64 actions, eight classes, $\delta = 2$). Class recovery is measured over 10 seeds and candidate reduction over 20 seeds. Motor clustering is K-means on motor descriptors, given the true number of classes and the same candidate budget as EAQ.

	Class recovery		Candidate reduction	
Method	ARI ↑	Exact	Candidates ↓	Norm. regret ↓
All actions	–	–	64	0
Motor clustering	0.009	0/10	8	0.331
EAQ (complete linkage)	1.000	10/10	8	0

Table 3: Online EAQ versus Static on the six-episode HM3D pilot suite (180 decisions). Metrics are averaged over decisions.

Metric	Static	Online EAQ
ARI ↑	0.285	0.337
PPE ↓	0.397	0.345
FMP ↓	0.269	0.242
False merges per decision ↓	5.66	5.08
False splits per decision ↓	2.68	2.17
Mean number of classes	2.44	2.44

(Table 1). In every seed, the task-conditioned partition differs from the motor partition, and the q_{C} and q_{L} partitions differ from each other. It also differs from a task-agnostic partition that averages over all sectors in 19 of 20 seeds for q_{C} and 8 of 20 for q_{L}. All partitions are unchanged between 256 and 512 pixel resolution.

Motor geometry as a proxy. In a synthetic construction with 64 actions, we interpolate motor descriptors from independent of the consequence classes ($\lambda = 0$) to aligned with them ($\lambda = 1$). The absolute Spearman correlation between motor and epistemic distances rises from 0.041 to 0.998, and 7-nearest-neighbor disagreement falls from 0.940 to 0 (Figure 2). Motor similarity can thus range from uninformative to fully aligned with epistemic identity, and cannot replace a consequence-based criterion in general.

These results answer Q1: action identity in the tested embodied settings is compressive, changes with the demand, and does not follow motor geometry. Demand changes are therefore mandatory reconsideration events, but the frequent within-demand changes also show that a fixed demand alone does not justify reuse; fixed-demand reuse requires a stability condition.

5.2 Recovery with Known Ground Truth

Because the partitions in Section 5.1 have no ground truth, we test complete linkage in synthetic constructions with 64 actions in eight well-separated Gaussian consequence classes (Sections F and H; Table 2). At $\delta = 2$, complete linkage on empirical Wasserstein distances recovers the generating classes exactly in all 10 seeds, and at every tested threshold from 0.75 to 4, whereas K-means on motor descriptors performs at chance level even when given the true number of classes. Evaluating one representative per block reduces 64 candidate evaluations to 8 with no observed regret in any of 20 seeds, whereas motor clustering with the same budget incurs a mean normalized regret of 0.331. In two further studies (Sections I and K), a single refresh after a demand switch raises ARI from -0.024 to 1 in all 10 seeds, and the partition remains unchanged under calibrated perturbations of the ground metric. When classes are well separated, complete linkage thus behaves as the theory predicts, and grouping by consequence preserves decision quality where grouping by motor similarity does not (Q2).

Table 4: Refresh/consensus ablation on the six-episode HM3D pilot. All variants use the same seed-8603 initialization. Lower PPE/FMP/FM/FS and higher ARI are better.

Method	PPE	FMP	ARI	FM	FS	K
Static	0.3971	0.2693	0.2848	5.656	2.683	2.444
Refresh-single	**0.3452**	**0.2381**	0.3320	**5.000**	2.250	2.500
Refresh-mean	0.3474	0.2616	**0.3671**	5.494	**1.800**	2.333
Refresh-median	0.3495	0.2468	0.3318	5.183	2.156	2.389
Online EAQ	**0.3452**	0.2421	0.3368	5.083	2.167	2.444

5.3 Online EAQ in HM3D

The Online EAQ pilot uses six HM3D episodes from three scenes (180 decisions, 12 demand switches), disjoint from training and Q1 evaluation. Against a Static baseline with identical initialization, we report pairwise partition error (PPE), false-merge proportion (FMP), and ARI; the counterfactual reference partition is computed only after both methods run (Section L).

Online EAQ raises ARI from 0.285 to 0.337 and lowers PPE from 0.397 to 0.345 at the same mean number of classes (Table 3). Four of 12 refreshes change the partition. Agreement is lowest in scene S3 (ARI < 0.08), where shared prediction errors dominate, suggesting that consequence prediction rather than refresh timing is the main remaining limitation.

What causes the online gain? We separate task-triggered refresh from the three-predictor consensus while keeping the seed-8603 initialization fixed. Table 4 shows that task-triggered refresh is the main source of improvement: Refresh-single reduces PPE from 0.3971 to 0.3452 and raises ARI from 0.2848 to 0.3320, already recovering most of the full Online EAQ gain (PPE 0.3452, ARI 0.3368). The ensemble primarily changes the error trade-off. In particular, a simple mean ensemble attains the highest ARI (0.3671) and the fewest false splits, but has a higher false-merge proportion (0.2616) than the hysteretic rule (0.2421). False merges and false splits have asymmetric consequences for action abstraction: a false split primarily sacrifices compression, whereas a false merge can violate the within-block substitutability premise and cause a planner to discard a non-interchangeable action. We therefore do not select the consensus rule by ARI alone. The asymmetric rule deliberately trades some overall partition agreement for a lower false-merge rate and should be viewed as a conservative robustness layer rather than the principal cause of the online improvement.

Does the stability condition identify safe reuse? We next connect Theorem 3.7 to the observed HM3D dynamics. After excluding demand switches and episode boundaries, the pilot contains 162 consecutive fixed-demand transitions, of which 93 (57.4%) change the realized-branch reference partition. For each transition we compute the current separation margin ρ_t and two drift diagnostics. The first is the realized-branch point-mass analogue of the theorem's actionwise drift, yielding the condition $2\eta_t < \rho_t$. It certifies 25 transitions, and none of the 25 changes partition. The second uses the reference pairwise-distance drift computed from counterfactual reference branches

$$B_t = \max_{a,a'} |D_{t+1}(a, a') - D_t(a, a')| \,. \tag{14}$$

The condition $B_t < \rho_t$ certifies 28 transitions, again with zero observed changes; among the remaining 134 transitions, 93 (69.4%) change partition. Certificate failure is not evidence that a change must occur: both conditions are conservative sufficient tests. The result instead supplies an empirical link between the stability theorem and online maintenance: when the realized-branch analogue of the theorem's sufficient stability condition is satisfied, the reference partition is unchanged on every certified transition in this pilot.

Together, the ablation and stability diagnostic sharpen Q3. Task-triggered refresh provides most of the measured gain over static caching, while the realized-branch diagnostics identify a conservative subset of within-demand transitions that is empirically consistent with the theorem's sufficient reuse condition. A deployable drift-triggered extension would require estimating an upper bound on η_t or B_t from runtime information rather than from counterfactual reference branches.

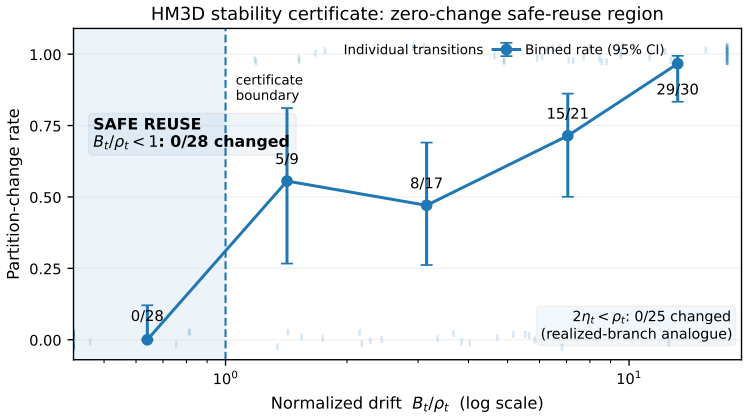


Figure 3: **The stability condition identifies a conservative safe-reuse region in HM3D.** Each mark denotes a fixed-demand transition and the connected markers show the observed partition-change rate in bins of normalized pairwise-distance drift B_t/ρ_t; error bars are two-sided 95% Wilson intervals. In the certificate-safe region $B_t/\rho_t < 1$, none of 28 transitions changes the reference partition. Beyond the certificate boundary, partition changes are common, reaching 29/30 transitions in the largest-drift bin. The theorem-aligned realized-branch condition $2\eta_t < \rho_t$ independently certifies 25 transitions, again with zero observed changes. Certificate failure is not a prediction of change; the conditions are sufficient rather than necessary.

6 Conclusion

EAQ defines the identity of sensing actions by their task-conditioned epistemic consequences rather than by actuation. Exact equivalence preserves value for every consequence-only utility, bounded-diameter partitions bound within-block regret, and well-separated partitions are stable under small drift. In the tested embodied settings, actions that differ in actuation repeatedly share a class and the classes often shift with the demand. In the HM3D pilot, task-triggered refresh accounts for most of the improvement over static caching, while the three-predictor consensus mainly adjusts the false-merge/false-split trade-off. The theory–experiment connection is deliberately scoped: all 25 fixed-demand transitions satisfying the realized-branch analogue of $2\eta_t < \rho_t$, and all 28 satisfying the pairwise diagnostic $B_t < \rho_t$, preserve the reference partition, which is empirically consistent with the theorem's sufficient stability condition but is not a direct validation of its population-law premise. These diagnostics are sufficient rather than necessary and are evaluated from counterfactual references, so a deployable drift-triggered monitor remains open. The pilot is also small, and candidate reduction is evaluated only in synthetic settings; downstream embodied planning and broader-scene evaluation are natural next steps.

Generative AI tools, including ChatGPT, Codex, and Claude, assisted with feedback on the conceptual framework, experiment design, software implementation, analysis scripts, manuscript drafting, and figure preparation. The authors verified all reported results against the experimental records and take full responsibility for the manuscript.

Reproducibility Statement

The appendix supplies assumptions, proofs, controlled-study settings and results, and full-precision pilot metrics. Code, configurations, trained checkpoints, and the pilot-suite manifest will be included in the anonymous supplementary material.

References

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923. PMLR, 2016. URL https://proceedings.mlr.press/v48/abel16.html.

David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1639–1650. PMLR, 2020.

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. doi: 10.1007/BF00133571.

Mahsa Asadi, Mohammad Sadegh Talebi, Hippolyte Bourel, and Odalric-Ambrym Maillard. Model-based reinforcement learning exploiting state-action equivalence. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pp. 204–219. PMLR, 2019. URL https://proceedings.mlr.press/v101/asadi19a.html.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. doi: 10.1109/5.5968.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, 1951.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 941–950. PMLR, 2019. URL https://proceedings.mlr.press/v97/chandak19a.html.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X.

Mahsa Ghasemi, Erdem Bulgur, and Ufuk Topcu. Task-oriented active perception and planning in environments with partially known semantics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3484–3493. PMLR, 2020. URL https://proceedings.mlr.press/v119/ghasemi20a.html.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Pengjie Gu, Mengchen Zhao, Chen Chen, Dong Li, Jianye Hao, and Bo An. Learning pseudometric-based action representations for offline reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7902–7918. PMLR, 2022. URL https://proceedings.mlr.press/v162/gu22b.html.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3477–3484, 2016. doi: 10.1109/ICRA.2016.7487527.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Siddharth Karamcheti, Megha Srivastava, Percy Liang, and Dorsa Sadigh. LILA: Language-informed latent actions. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1379–1390. PMLR, 2022. URL https://proceedings.mlr.press/v164/karamcheti22a.html.

Yining Li, Peizhong Ju, and Ness Shroff. Achieving sample and computational efficient reinforcement learning by action space reduction via grouping. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5afaa8b4dd18eb1eed055d2d821b58ae-Abstract-Conference.html.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Balaraman Ravindran. *An Algebraic Approach to Abstraction in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2004. URL https://people.cs.umass.edu/~barto/courses/cs687/RavindranThesis.pdf.

Guy Tennenholtz and Shie Mannor. The natural language of actions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6196–6205. PMLR, 2019.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45d74e190008c7bff2845ffc8e3facd3-Abstract-Conference.html.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

A Appendix

This appendix makes explicit the measure-theoretic and finite-action conditions used in the main text. At a fixed pre-action epistemic condition e_t and task q_t, let $P_{a,t}^{M_t} = \mathrm{Law}(Y_a^{q_t} \mid e_t)$ for $M_t = M_{q_t}$. The condition e_t is the information locally available before acting. Let $\mathcal{A}$ be a finite, nonempty candidate action set. At time t, $(\mathcal{E}_{M_t}, d_{M_t})$ is a Polish metric space and every task-conditioned consequence law $P_{a,t}^{M_t}$ belongs to $\mathcal{P}_1(\mathcal{E}_{M_t})$. Hence $W_{1,d_{M_t}}$ is finite. For HM3D update vectors the declared ground metric is half-L_1; the other controlled constructions retain their stated metrics. A consequence utility $U : \mathcal{E}_{M_t} \to \mathbb{R}$ is measurable and integrable under the laws being compared.

Whenever a regret certificate is invoked, U is also L-Lipschitz with respect to d_{M_t}. Finiteness of $\mathcal{A}$ gives existence of every displayed maximum and minimum over actions or pairs.

The laws may arise from a measurable map $M_t : \mathcal{E} \to \mathcal{E}_{M_t}$ applied to the unprojected post-action effect law $P^{\mathrm{eff}}_{a|e_t}$, conditioned on the same pre-action e_t: $P^{M_t}_{a,t} = (M_t)_{\#} P^{\mathrm{eff}}_{a|e_t}$. This representation is illustrative rather than restrictive; the results use only the task-conditioned laws and declared metrics. Strong separation in Assumption 3.6 is not part of the exact quotient definition and is not needed for Theorem 3.5.

For cross-time statements, suppose there is a common Polish metric space $(\bar{\mathcal{E}}, \bar{d})$ and isometric embeddings $\iota_t : \mathcal{E}_{M_t} \to \bar{\mathcal{E}}$ and $\iota_{t+1} : \mathcal{E}_{M_{t+1}} \to \bar{\mathcal{E}}$. Define

$$Q_{a,s} = (\iota_s)_{\#} P^{M_s}_{a,s}, \qquad s \in \{t, t+1\}. \tag{15}$$

Isometry gives $D_s(a, a') = W_{1,\bar{d}}(Q_{a,s}, Q_{a',s})$. Without a common comparable outcome space, the temporal stability statement is not asserted.

A.1 Complete-linkage recovery under input separation

We verify the sufficient condition in Equation (6) for the exact deterministic procedure in Definition 3.4. Let $\theta = \delta + \epsilon_{\mathrm{alg}}$. Initially every current block lies within one target block of $\mathcal{T}$. Suppose this remains true after some merges. The complete-linkage distance between current blocks lying in distinct target blocks is strictly greater than θ, so such a merge is ineligible. If a target block is represented by at least two current blocks, their complete-linkage distance is at most θ, so at least one within-target merge remains eligible. The selected minimum-distance pair must therefore be a within-target pair until each target block has been assembled. At that point every remaining pair crosses target blocks and is ineligible, so the procedure stops with exactly $\mathcal{T}$.

This induction also covers the boundary convention: a within-target distance equal to θ remains eligible, while the strict cross-target inequality excludes equality there. Lexicographic tie handling makes the output deterministic when several pairs have the same distance; under the separation condition it cannot change the recovered target. With $G = D_M$ and $\epsilon_{\mathrm{alg}} = 0$, this proves the population-distance recovery statement used in Theorem 3.7.

B Proofs of Exact Value Equivalence and the Regret Bound

B.1 Exact epistemic value equivalence

Proof of Proposition 3.2. By Definition 3.1, $a \equiv_{e,M} a'$ means $P^M_a = P^M_{a'}$ as probability measures on $(\mathcal{E}_M, \mathcal{B}(\mathcal{E}_M))$. Because U is measurable and integrable under this common law,

$$\mathbb{E}_{P^M_a}[U] = \int_{\mathcal{E}_M} U(y)\, P^M_a(\mathrm{d}y) = \int_{\mathcal{E}_M} U(y)\, P^M_{a'}(\mathrm{d}y) = \mathbb{E}_{P^M_{a'}}[U].$$

The statement concerns only utilities of the task-conditioned epistemic consequence. It makes no assertion about variables excluded from that consequence space. □

B.2 Population-diameter cluster regret

Proof of Theorem 3.5. Let $a, a' \in \tau$. The Kantorovich–Rubinstein dual inequality on the Polish space $(\mathcal{E}_M, d_M)$ gives

$$\left| \int U \,\mathrm{d}P^M_a - \int U \,\mathrm{d}P^M_{a'} \right| \le L W_{1,d_M}(P^M_a, P^M_{a'}). \tag{16}$$

The required integrals are finite by assumption (and also follow from finite first moments for any finite-valued Lipschitz U). Since $\mathcal{A}$ is finite, $a^\star_\tau$ exists. Substituting $a = a^\star_\tau$ and $a' = \tilde{a}$ into Equation (16) yields

$$V_M(a^\star_\tau) - V_M(\tilde{a}) \le L W_{1,d_M}(P^M_{a^\star_\tau}, P^M_{\tilde{a}}) \le L \operatorname{diam}_M(\tau).$$

For a block of any population-δ-feasible partition, $\operatorname{diam}_M(\tau) \le \delta$, completing the proof. No minimum-cardinality property is used. □

Corollary B.1 (Ground-metric misspecification)**.** *Fix the task-conditioning map and consequence laws. Let $d_M^\star$ be a task-relevant ground metric and $\widetilde{d}_M$ the metric used operationally. Suppose that for all consequence-law pairs under consideration,*

$$\left|W_{1,d_M^\star}(P,Q) - W_{1,\widetilde{d}_M}(P,Q)\right| \le \varepsilon_M. \tag{17}$$

If $\operatorname{diam}_{\widetilde{d}_M}(\tau) \le \delta$, then $\operatorname{diam}_{d_M^\star}(\tau) \le \delta + \varepsilon_M$. If U is L-Lipschitz under $d_M^\star$, every $\tilde{a} \in \tau$ has regret at most $L(\delta + \varepsilon_M)$ relative to the best action in τ.

Proof. For every $a, a' \in \tau$, Equation (17) implies

$$W_{1,d_M^\star}(P_a^M, P_{a'}^M) \le W_{1,\widetilde{d}_M}(P_a^M, P_{a'}^M) + \varepsilon_M \le \delta + \varepsilon_M.$$

Taking the maximum over the finite block gives the diameter result. Applying Theorem 3.5 with $d_M^\star$ gives the regret result. □

This corollary does not cover misspecification of M or changes in the consequence laws. A sufficient, stronger premise for Equation (17) is uniform ground-metric distortion $\sup_{y,y'} |d_M^\star(y,y') - \widetilde{d}_M(y,y')| \le \varepsilon_M$ on the common support.

C Proof of Ideal and Population Complete-Linkage Stability

Lemma C.1 (Pairwise distance drift)**.** *Under Equation* (10)*, every $a, a' \in \mathcal{A}$ satisfies*

$$|D_{t+1}(a,a') - D_t(a,a')| \le 2\eta_t. \tag{18}$$

Proof. For any four probability measures in a common Wasserstein space, the triangle inequality gives

$$W_1(Q_{a,t+1}, Q_{a',t+1}) \le W_1(Q_{a,t+1}, Q_{a,t}) + W_1(Q_{a,t}, Q_{a',t}) + W_1(Q_{a',t}, Q_{a',t+1}).$$

Subtracting $D_t(a,a')$ and applying Equation (10) bounds the increase by $2\eta_t$. Interchanging t and $t+1$ bounds the decrease by the same quantity, proving the absolute bound. □

Proof of Theorem 3.7. Let $2\eta_t < \rho_t$. At time t, Equations (8) and (9) make $\mathcal{T}_t$ population-δ-feasible and prevent any feasible block from containing actions from two different blocks of $\mathcal{T}_t$. Hence every feasible partition refines $\mathcal{T}_t$, while any strict refinement has more blocks. Thus $\mathcal{T}_t$ is the unique ideal solution $\mathcal{P}_t^{\star,\delta}$. Applying Equation (6) to $G = D_t$ and $\epsilon_{\text{alg}} = 0$ also gives $\mathcal{P}_{t,\text{CL}}^{\delta} = \mathcal{T}_t$.

If a, a' lie in the same block $\tau \in \mathcal{T}_t$, Equations (8) and (18) imply

$$D_{t+1}(a,a') \le D_t(a,a') + 2\eta_t \le \delta - \rho_t + 2\eta_t < \delta.$$

Thus every target block remains diameter-feasible at $t+1$. If $a \in \tau_i$ and $a' \in \tau_j$ for distinct target blocks, Equations (9) and (18) imply

$$D_{t+1}(a,a') \ge D_t(a,a') - 2\eta_t \ge \delta + \rho_t - 2\eta_t > \delta.$$

No feasible block at $t+1$ can therefore contain actions from two different target blocks. Every feasible partition at $t+1$ is a refinement of $\mathcal{T}_t$, whereas $\mathcal{T}_t$ itself remains feasible. Any strict refinement has more blocks because $\mathcal{A}$ is finite and nonempty. Hence $\mathcal{P}_{t+1}^{\star,\delta} = \mathcal{T}_t$ uniquely. The same within-at-most-δ and cross-strictly-above-δ inequalities satisfy Equation (6) for $G = D_{t+1}$ and $\epsilon_{\text{alg}} = 0$, giving $\mathcal{P}_{t+1,\text{CL}}^{\delta} = \mathcal{T}_t$. Combining the four equalities proves Equation (11). The complete-linkage conclusion is independent of tie order under this separation, although the implementation retains its deterministic lexicographic rule. □

D Conditional Results for Predicted Distances

The two propositions concern different prediction objects. Proposition D.1 predicts consequence *laws at the next decision time* $t+1$ from time t, with one predictor, a known uniform matrix-error bound, and an ideal minimum-cardinality objective. Online EAQ instead predicts candidate actions' *post-action belief points at the current authorized refresh decision*, combines three checkpoints through a cache-dependent max/median rule, and runs complete linkage. Although time advances while an action executes, these current-decision predictions are not the next-decision laws of Proposition D.1. Proposition D.2 addresses the authorized refresh calculation, conditional on population-error and separation premises.

Suppose a predictor at time t outputs next-time task-conditioned consequence laws $\widehat{P}^{M_{t+1}}_{a,t+1|t}$ in $\mathcal{P}_1(\mathcal{E}_{M_{t+1}})$. Define its pairwise distance matrix

$$\widehat{D}_{t+1|t}(a,a') = W_{1,d_{M_{t+1}}}\left(\widehat{P}^{M_{t+1}}_{a,t+1|t}, \widehat{P}^{M_{t+1}}_{a',t+1|t}\right) \tag{19}$$

and the matrix approximation error

$$\omega_{t+1|t} = \max_{a,a'\in\mathcal{A}} \left|\widehat{D}_{t+1|t}(a,a') - D_{t+1}(a,a')\right|. \tag{20}$$

Let $\widehat{\mathcal{P}}^{\star,\delta}_{t+1}$ be an ideal minimum-cardinality partition under $\widehat{D}_{t+1|t}$, using the same diameter threshold and optimization rule as Definition 3.3. This conditional object is not the empirical complete-linkage estimator in Definition 3.4.

Empirical validation is a separate object. If observations at $t+1$ yield empirical laws $\widetilde{P}^{M_{t+1}}_{a,t+1}$, define

$$D^{\mathrm{emp}}_{t+1}(a,a') = W_{1,d_{M_{t+1}}}\left(\widetilde{P}^{M_{t+1}}_{a,t+1}, \widetilde{P}^{M_{t+1}}_{a',t+1}\right), \quad \nu^{\mathrm{emp}}_{t+1} = \max_{a,a'} |D^{\mathrm{emp}}_{t+1}(a,a') - D_{t+1}(a,a')|. \tag{21}$$

No sum of $\omega_{t+1|t}$ and ν^{emp}_{t+1} is used: they describe distinct matrices unless an algorithm explicitly combines them.

Proposition D.1 (Conditional learned tracking)**.** *Assume the finite-action setup, Assumption 3.6 and Equation* (10)*, and the error definition in Equation* (20)*. If*

$$2\eta_t + \omega_{t+1|t} < \rho_t, \tag{22}$$

then the true time-t and time-$(t+1)$ ideal partitions and the predicted ideal partition are unique and

$$\widehat{\mathcal{P}}^{\star,\delta}_{t+1} = \mathcal{P}^{\star,\delta}_{t+1} = \mathcal{P}^{\star,\delta}_{t} = \mathcal{T}_t. \tag{23}$$

Proof. Condition Equation (22) implies $2\eta_t < \rho_t$, so Theorem 3.7 gives $\mathcal{P}^{\star,\delta}_{t+1} = \mathcal{P}^{\star,\delta}_{t} = \mathcal{T}_t$. For a, a' in one target block,

$$\widehat{D}_{t+1|t}(a,a') \le D_{t+1}(a,a') + \omega_{t+1|t} \le \delta - \rho_t + 2\eta_t + \omega_{t+1|t} < \delta.$$

For actions in distinct target blocks,

$$\widehat{D}_{t+1|t}(a,a') \ge D_{t+1}(a,a') - \omega_{t+1|t} \ge \delta + \rho_t - 2\eta_t - \omega_{t+1|t} > \delta.$$

Thus the target blocks are feasible under the predicted matrix and no predicted-feasible block can cross them. As in the proof of Theorem 3.7, every predicted-feasible partition is a refinement of the target partition, which is therefore the unique minimum-cardinality solution. This proves Equation (23). □

Applying the proposition requires a bound on $\omega_{t+1|t}$, which Online EAQ does not estimate.

D.1 Conditional authorized three-checkpoint refresh

The following statement applies to one current decision at which a task-cache update is authorized. It assumes errors against the unknown population-law distance.

Proposition D.2 (Conditional consensus-refresh bounds)**.** *Fix pre-action condition e, task q, and input task cache C. Let $D_{e,q}(a,a') = W_{1,d_{M_q}}(P^q_{a|e}, P^q_{a'|e})$ be the true population distance and $\theta = \delta + \epsilon_{\mathrm{alg}}$. For the three validated symmetric finite current-decision predictor matrices $\widehat{d}_j$, set*

$$\omega_j = \max_{a \neq a'} |\widehat{d}_j(a,a') - D_{e,q}(a,a')|, \qquad \omega_{(1)} \leq \omega_{(2)} \leq \omega_{(3)} \tag{24}$$

to be their sorted uniform pair-matrix errors. For a pair separated in C, Equation (13) *satisfies*

$$D_{e,q} - \omega_{(1)} \leq \widetilde{D} \leq D_{e,q} + \omega_{(3)}; \tag{25}$$

for a pair merged in C, it satisfies

$$D_{e,q} - \omega_{(2)} \leq \widetilde{D} \leq D_{e,q} + \omega_{(2)}. \tag{26}$$

Every block of $\mathrm{CL}_{\delta,\epsilon_{\mathrm{alg}}}(\widetilde{D})$ has true population-distance diameter at most $\theta + \omega_{(2)}$. Moreover, if a target partition T has every within-block true pair distance at most $\theta - \gamma$ and every cross-block distance at least $\theta + \gamma$, where $\gamma > 0$ and $\omega_{(3)} < \gamma$, then this complete-linkage output equals T. For the $\delta \pm \rho$ separation convention in Assumption 3.6, $\omega_{(3)} + \epsilon_{\mathrm{alg}} < \rho$ is sufficient.

Proof. For a cached-separated pair, choose the model with smallest uniform error. Its predicted distance is at least $D_{e,q} - \omega_{(1)}$, hence so is the maximum; every model is at most $D_{e,q} + \omega_{(3)}$. For a cached-merged pair, at least two models lie in $[D_{e,q} - \omega_{(2)}, D_{e,q} + \omega_{(2)}]$, so their median lies in that interval. Complete linkage merges blocks only if every crossing input distance is at most θ. Induction over its merges shows that every pair in a final block has $\widetilde{D} \leq \theta$. For a cached-separated pair this and the smallest-error model give $D_{e,q} \leq \theta + \omega_{(1)}$; for a cached-merged pair the median bound gives $D_{e,q} \leq \theta + \omega_{(2)}$. Thus the claimed block-diameter bound holds. Under the target separation and $\omega_{(3)} < \gamma$, both cache-dependent cases have input distance strictly below θ within T and strictly above θ across T. The complete-linkage recovery induction in Section A.1 therefore yields T. Under $\delta \pm \rho$ separation, the stated condition accounts for the offset $\theta = \delta + \epsilon_{\mathrm{alg}}$ and implies the same strict input separation. □

The two-small-error-model bound gives a one-sided true diameter limit, not full target recovery: the maximum for a cached-separated pair can be inflated by the third model and block a valid merge. Applying the proposition requires the population-level errors ω_j, which agreement among checkpoints does not by itself provide, and it concerns authorized refreshes rather than the single-model initialization.

E Common Evaluation and Statistical Conventions

The six controlled studies (the separated-class recovery, motor–epistemic geometry, candidate-reduction, demand-switch refresh, controlled Habitat, and metric-corruption studies; Appendices F–K) use configurations, metrics, thresholds, seed lists, and decision rules specified before their reported runs. The seed is the inferential unit. Multiple tasks, action pairs, windows, resolutions, or arrays within a seed are aggregated or treated descriptively, not counted as independent repetitions. Unless otherwise noted, continuous summaries report the mean, sample standard deviation with $\mathrm{ddof} = 1$, and an untruncated two-sided Student-t 95% confidence interval across seeds. Exact counts are reported as counts; post hoc intervals are not attached to every count.

Adjusted Rand index (ARI) compares a recovered partition with generating labels: 1 is exact agreement, 0 is its chance-adjusted baseline, and negative values are possible. False-merge rate (FMR) and false-split rate (FSR) record pairwise class errors where structurally applicable; lower is better. Pair disagreement (PD) is the fraction of action pairs whose co-membership differs between two partitions; $\mathrm{PD} = 0$ means identical pairwise membership. The k-nearest-neighbor disagreement (KND) is the mean fraction of nonshared neighbors and is likewise minimized at 0. Normalized regret (NR) is calculated by each evaluation's specified utility normalization; lower is better.

The empirical partitions in these controlled studies are produced by complete linkage over each evaluation's pairwise distance matrix. Complete linkage enforces a diameter threshold but need not give a globally minimum-cardinality partition. Online DEAQ uses a separate pilot suite, described in

Appendix L. The ablation and stability diagnostic of Section 5.3 and **??** were specified after the pilot results and are not among these pre-specified studies.

Across these six controlled studies, observations or synthetic consequence samples define action consequences; a fixed task map retains the demand-relevant component; the declared ground metric induces pairwise distances; and deterministic complete linkage produces a thresholded partition. Where decision quality is evaluated, one task-independent representative per block becomes visible to the planner. Population labels and values are evaluation references, not clustering inputs.

For the recovery, geometry, candidate-reduction, and corruption studies, a seed indexes a newly generated synthetic world and its named sampling streams; the geometry study additionally generates its independent motor-noise component. The refresh study uses a seed-specific collection of observation-noise slices shared by the methods compared within that seed. In the Habitat study, a seed indexes one sampled pose and the complete paired acquisition from that pose. Cross-seed intervals summarize variation over these randomized instances.

F Separated-Class Recovery Study: Setting and Results

Setting. The separated-class recovery study is a strongly separated balanced Gaussian construction with 64 actions, eight generating consequence classes, and 100 observations per action in the primary setting. Every action in class $c(a)$ has consequence law $\mathcal{N}(\mu_{c(a)}, 0.5^2 I_4)$. Independently drawn raw class centers are centered and uniformly scaled so that their minimum Euclidean separation is 4; the six-dimensional motor descriptors are sampled independently of class. The distance is exact discrete Wasserstein-1 between equally weighted empirical measures with Euclidean ground cost, the primary threshold is $\delta = 2$, and the evaluation seeds are 100–109 ($n = 10$). Oracle labels provide a sanity check. Motor K-means is a constructed negative control because motor descriptors are class-independent in this design.

Tables 5, 6, and 7, together with Fig. 4, report the primary, sample-size, and threshold-grid views.

Table 5: Recovery study: primary seed-level results at $\delta = 2$. Intervals are two-sided Student-t 95% intervals over ten seeds.

Method	ARI [95% CI]	FMR [95% CI]	FSR [95% CI]	Mean blocks	Exact
Oracle	1.000 [1,1]	0 [0,0]	0 [0,0]	8	10/10
DEAQ empirical W_1	1.000 [1,1]	0 [0,0]	0 [0,0]	8	10/10
Motor K-means	0.009 [−0.006,0.023]	0.131 [0.125,0.138]	0.859 [0.843,0.876]	8	0/10

Table 6: Recovery study: sample-size sensitivity at $\delta = 2$. Values are means over ten seeds; exact gives the number of seeds matching the generating labels.

Samples/action	1	2	5	10	25	50	100
Mean ARI	.733	.946	1	1	1	1	1
Mean blocks	16.1	9.6	8	8	8	8	8
Exact	0/10	1/10	10/10	10/10	10/10	10/10	10/10

DEAQ's pairwise AUROC was 1 [1,1]. Its paired ARI advantage over Motor K-means was 0.9913 with 95% CI [0.9768, 1.0059], and DEAQ had higher ARI in 10/10 seeds. The upper confidence limit above one belongs to the untruncated interval for a paired difference, not an ARI estimate.

At one sample per action, AUROC was approximately one (mean 0.99996) without complete partition recovery, illustrating that pair ranking and thresholded partition recovery are different targets.

Table 7: Recovery study: threshold sensitivity. Each entry is the mean over ten seeds.

δ	ARI	FMR	FSR	Blocks
0	0	0	1	64
0.25	0	0	1	64
0.5	0.331	0	0.780	33.3
0.75	1	0	0	8
1	1	0	0	8
1.5	1	0	0	8
2	1	0	0	8
2.5	1	0	0	8
3	1	0	0	8
3.5	1	0	0	8
4	1	0	0	8
5	0.791	0.057	0	6.5
6	0.687	0.096	0	5.8

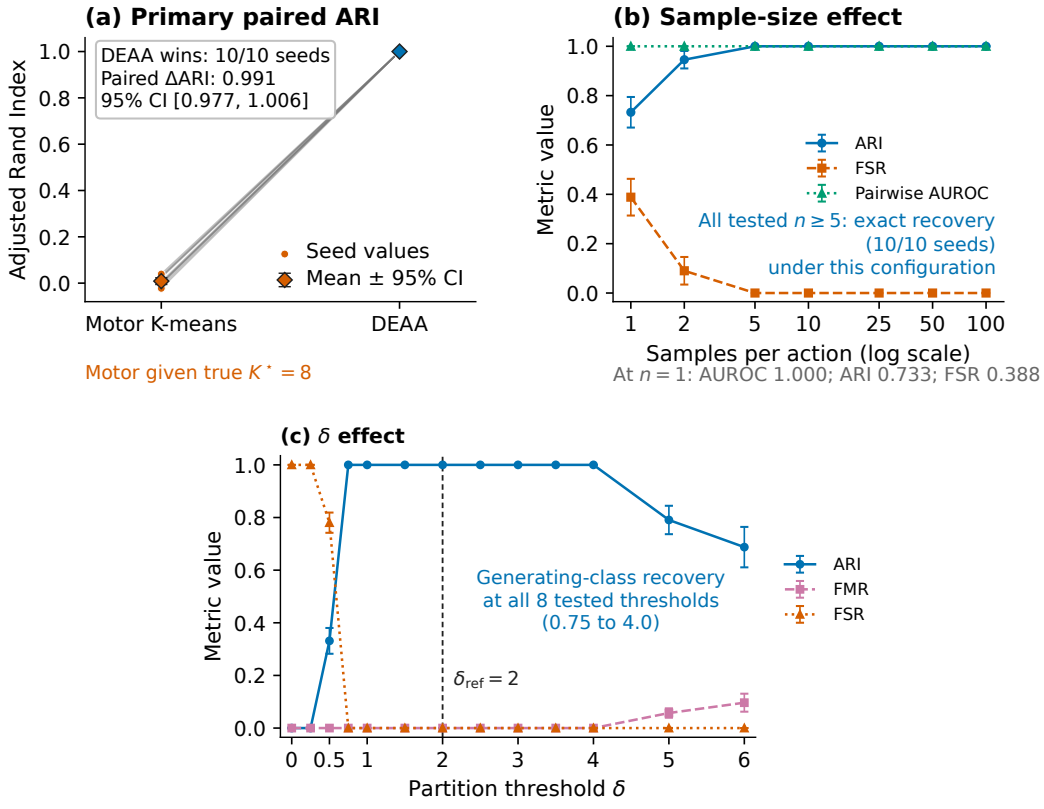


Figure 4: Recovery study: primary recovery, sample-size sensitivity, and the complete threshold grid. The estimator matches the generating labels at every tested threshold from 0.75 through 4.

G Motor–Epistemic Geometry Study: Setting and Results

Setting. The motor–epistemic geometry study uses the same 64-action, eight-class controlled consequence construction and seeds 200–209. A coupling parameter $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ interpolates from class-independent to class-center-informed motor descriptors. Spearman correlation compares the 2,016 motor and epistemic pairwise distances within each seed. KND measures disagreement of k-nearest-neighbor sets; $k = 7$ is primary. Pairwise values are reduced to one estimate per seed before inference.

Specifically, the four-dimensional class center is zero-padded to $z_i \in \mathbb{R}^6$, independent noise is $\epsilon_i \sim \mathcal{N}(0, I_6)$, and s_z, s_ϵ are the respective medians of all pairwise Euclidean distances. The motor descriptor is

$$m_i(\lambda) = \lambda z_i / s_z - \sqrt{1 - \lambda^2}\, \epsilon_i / s_\epsilon.$$

Motor distance is Euclidean distance between these descriptors. Epistemic distance is exact discrete W_1 with Euclidean ground cost between the equally weighted empirical consequence measures; the same epistemic matrix is reused for every λ. Thus $\lambda = 1$ aligns motor distance with population center geometry, not necessarily with its finite-sample empirical estimate.

Tables 8 and 9, together with Fig. 2, report the endpoint and coupling-trajectory results.

Table 8: Geometry study: primary endpoint results over ten seeds.

Metric	Mean	Sample SD	95% CI
$\lvert\rho(0)\rvert$	0.0408	0.0434	[0.0098,0.0718]
KND(0), $k = 7$	0.9400	0.0054	[0.9361,0.9438]
$\rho(1)$	0.9978	0.0006	[0.9974,0.9982]
KND(1), $k = 7$	0	0	[0,0]
$\Delta\rho$	1.0155	0.0582	[0.9739,1.0571]
ΔKND	0.9400	0.0054	[0.9361,0.9438]

Table 9: Geometry study: coupling trajectory. Values are seed means with 95% intervals.

λ	Spearman ρ	KND, $k = 7$
0	-0.0177 [-0.0593,0.0240]	0.9400 [0.9361,0.9438]
0.25	0.0678 [0.0417,0.0939]	0.9200 [0.9127,0.9272]
0.5	0.2997 [0.2765,0.3228]	0.8554 [0.8465,0.8643]
0.75	0.6527 [0.6193,0.6861]	0.6543 [0.6261,0.6824]
1	0.9978 [0.9974,0.9982]	0 [0,0]

Both paired effects were positive in 10/10 seeds. The endpoint $\lvert\rho(0)\rvert$ uses absolute correlation, whereas $\Delta\rho = \rho(1) - \rho(0)$ uses signed correlations.

H Candidate-Reduction Study: Setting and Results

Setting. The candidate-reduction study uses a strongly separated synthetic world with 64 actions, eight latent consequence classes, 100 samples per action, and seeds 300–319. Each action in class $c(a)$ induces $Y_a \sim \mathcal{N}(\mu_{c(a)}, 0.5^2 I_4)$; class centers have minimum separation 4. The eight tasks target the eight class centers, $g_{s,j} = \mu_{s,j}$, and are aggregated with equal weight within seed. Their utility and analytic population value are

$$U_{s,j}(y) = \exp\left[-\frac{\|y - g_{s,j}\|_2^2}{2\tau^2}\right], \qquad \tau = 2, \tag{27}$$

$$V_{s,j}(a) = \left(\frac{\tau^2}{\tau^2 + 0.5^2}\right)^2 \exp\left[-\frac{\|\mu_{c(a)} - g_{s,j}\|_2^2}{2(\tau^2 + 0.5^2)}\right]. \tag{28}$$

Raw action evaluation, an oracle partition, exact empirical-W_1 complete linkage, and standardized-descriptor Motor K-means are compared. Motor K-means receives the same seed-specific candidate

budget as DEAQ. The threshold grid is $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6\}$; $\delta = 2$ is primary. Six paired frontier contrasts use Bonferroni-adjusted intervals.

For the original motor descriptor $m_a \in \mathbb{R}^{d_m}$, the task-independent cost proxy is $C(a) = \|m_a\|_2^2/d_m$. Each block contributes the representative $r(B) = \arg\min^{\text{lex}}_{a\in B}(C(a), a)$ before any task value is read. The representative-only planner maximizes $(V_{s,j}(r(B)), -C(r(B)), -r(B))$ lexicographically over blocks. With $a^\star_{s,j}$ the full-set maximizer, raw regret is $R_{s,j} = V_{s,j}(a^\star_{s,j}) - V_{s,j}(a^{\mathcal{P}}_{s,j})$ and

$$\widetilde{R}_{s,j} = \frac{R_{s,j}}{V_{s,j}(a^\star_{s,j}) - \min_{a\in\mathcal{A}} V_{s,j}(a)}.$$

Candidate count includes only task values submitted to the planner; it excludes partition construction, distance estimation, and evaluation-only value access.

Table 10 reports the complete threshold trajectory.

Table 10: Candidate-reduction study: complete threshold trajectory. Candidates and NR are means over 20 seeds after within-seed task aggregation.

δ	DEAQ candidates	DEAQ NR	Motor candidates	Motor NR	DEAQ resource fraction
0	64.00	0	64.00	0	1.0000
0.25	64.00	0	64.00	0	1.0000
0.5	34.25	0	34.25	0	0.5352
0.75	8.00	0	8.00	0.3314	0.1250
1	8.00	0	8.00	0.3314	0.1250
1.5	8.00	0	8.00	0.3314	0.1250
2	8.00	0	8.00	0.3314	0.1250
2.5	8.00	0	8.00	0.3314	0.1250
3	8.00	0	8.00	0.3314	0.1250
3.5	8.00	0	8.00	0.3314	0.1250
4	8.00	0	8.00	0.3314	0.1250
5	6.70	0.1389	6.70	0.3996	0.1047
6	6.20	0.2004	6.20	0.4442	0.0969

At $\delta = 2$, Raw uses 64 candidates, while Oracle and the empirical deterministic complete-linkage DEAQ construction use 8. The latter's seed-mean NR is 0 [0,0], and all 160 within-seed task values have zero observed regret. Matched Motor mean NR is 0.3314; the paired Motor-minus-DEAQ interval is [0.2927,0.3701], positive in 20/20 seeds. At $\delta = 5$, its observed candidate count is 6.7, compression is $9.630\times$, and mean NR is 0.1389 with ordinary interval [0.1094,0.1684] and adjusted interval [0.0974,0.1804]. At $\delta = 6$, the corresponding values are 6.2, $10.537\times$, 0.2004, [0.1546,0.2463], and [0.1360,0.2649].

I Demand-Switch Refresh Study: Setting and Results

Setting. The demand-switch refresh study uses seeds 400–409 in a controlled synthetic system. The demand maps are $M_1 = \text{diag}(1, 1, 0, 0)$ and $M_2 = \text{diag}(0, 0, 1, 1)$; transformed four-dimensional samples use Euclidean ground cost and exact empirical W_1. One block applies an externally observed $M_1 \to M_2$ switch to the same underlying samples and compares the cached M_1 partition with one deterministic M_2 refresh; demand detection is not evaluated. A second block introduces one specified abrupt environment shift before window 5. Window 0 initializes the cache, windows 1–4 are stable, and windows 5–8 are post-shift monitoring windows.

At each monitoring window, deterministic complete linkage produces a canonical current partition. A partition different from the cache becomes a candidate with count 1; only recurrence of the same candidate in the next window raises the fixed persistence count to 2, triggers an alarm, and updates the cache at that boundary. Reversion to the cache clears the candidate, while a different candidate resets its count to 1. Detection and recovery delays are measured in window/update intervals as $t_{\text{alarm}} - 5$ and $t_{\text{recovery}} - 5$; a failure is retained explicitly and assigned the pre-specified capped value 4. The same seed drives all within-seed comparisons; windows are not independent repetitions.

Table 11 and Fig. 5 report the seed-level outcomes and the specified change trajectories.

Table 11: Refresh study: seed-level outcomes. Point intervals arise because all ten seed values coincide for these endpoints.

Block	Metric	Estimate or count	95% CI / status
Demand	Stale ARI	-0.0244	$[-0.0244,-0.0244]$
	Refreshed ARI	1	[1,1]
	Paired ARI improvement	1.0244	[1.0244,1.0244]
	Exact recovery	10/10	observed count
	Sentinel pairs $(0,8)$; $(0,16)$	10/10; 10/10	observed counts
Environment	Seed-level false-alarm rate	0	[0,0]
	Stable false alarms	0/40	descriptive windows
	Detection; recovery	10/10; 10/10	observed counts
	Capped detection; recovery delay	1; 1	[1,1]; [1,1]
	Final cached ARI	1	[1,1]
	Paired ARI improvement	0.2218	[0.2218,0.2218]

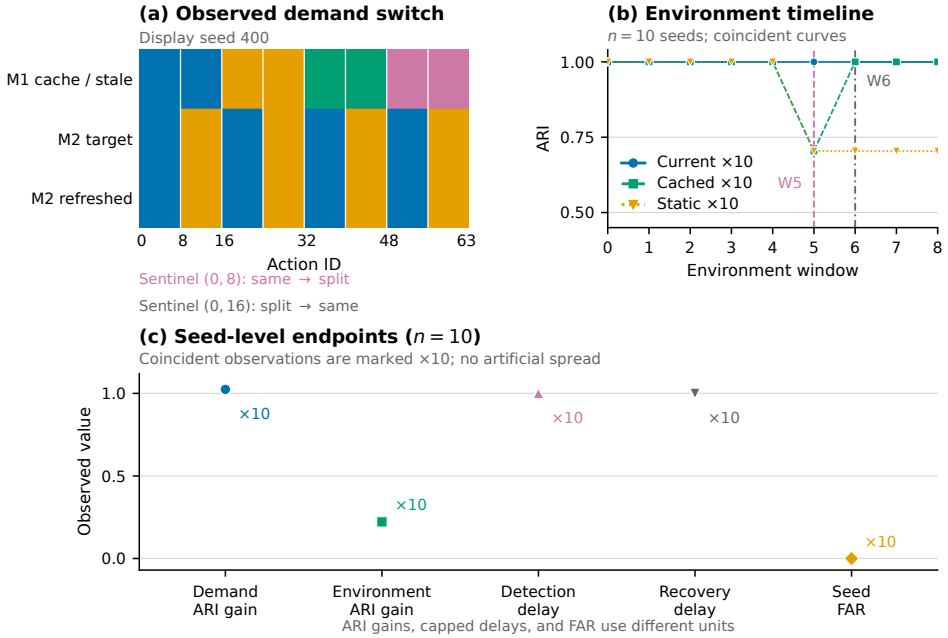


Figure 5: Refresh study: response to the observed demand switch and specified environment shift. In Panel A, colors denote canonical block identity and white dividers mark the eight-action classes. The ten seed trajectories in Panel B coincide and are shown without jitter or bands. Panel C shows ARI gains, capped delays, and FAR, which have different units.

J Controlled Habitat Study: Setting and Results

Setting. The controlled Habitat study uses one Skokloster Castle scene and navmesh, seven actions A00–A06, task queries q_C and q_L, sensor resolutions 256 and 512, and seeds 4401–4420. The actions are stationary RGB/depth acquisition without a simulator step (A00), sensor-node pan left/right by 10° (A01/A02), body translation left/right by 0.05 m (A03/A04), forward translation by 0.10 m (A05), and whole-body yaw left by 10° (A06). The primary pair P1 is A01 versus A06. Their motor vectors $(0,0,0,1)$ and $(0,0,1,0)$ have distance $\sqrt{2} > 0.50$, so the motor partition separates them. The seed-defined pose and complete acquisition are the sole inferential unit ($n = 20$); actions, pixels, tasks, and resolutions are within-seed measurements.

Each depth image is divided into left, central, and right sectors. Valid depth contributes to ten fixed bins and invalid depth to an eleventh category; a symmetric Dirichlet pseudocount $\alpha = 1$

gives posterior vector $b_{a,u}$ for action a and sector u. With uniform prior $b_{0,u}$, the consequence is $\Delta b_{a,u} = b_{a,u} - b_{0,u}$. Query q_{C} asks for forward/central directional clearance and retains $\Delta b_{a,C}$; q_{L} asks for left-sector clearance and retains $\Delta b_{a,L}$. For either query,

$$d_q(a, a') = \tfrac{1}{2}\|\Delta b_{a,q} - \Delta b_{a',q}\|_1 = \mathrm{TV}(b_{a,q}, b_{a',q}),$$

while the task-agnostic full distance averages the three sector-wise posterior TV distances. Deterministic complete linkage uses the primary threshold $\delta = 0.05$. Equivalence in this study therefore refers to co-membership in the operational partition at this threshold.

Table 1 reports the task-conditioned events, and Table 12 reports the resolution diagnostics.

Table 12: Habitat study: resolution diagnostics. The final three columns report denominator-state counts, not metric estimates.

Task	Exact 256/512	D mean $\pm$ SD [95% CI]	FMR app./N/A	FSR app./N/A	ARI reg./deg.
q_{C}	20/20	0 ± 0 [0,0]	17/3	20/0	17/3
q_{L}	20/20	0 ± 0 [0,0]	20/0	20/0	20/0

K Metric-Corruption Study: Setting and Results

Setting. The metric-corruption study uses seeds 500–519 and eight within-seed tasks under one deterministic, seed-normalized, population-calibrated anisotropic symmetric-positive-definite corruption family. Let

$$Q = \tfrac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \qquad h = (0, \tfrac{1}{3}, \tfrac{2}{3}, 1).$$

For population-center difference $v_{s,aa'}$ and $D_s = \max_{a<a'} \|v_{s,aa'}\|_2$, set $t_q = q\delta$, $\beta_{s,q} = t_q/(2D_s)$, $b_{s,q,k} = (1 - \beta_{s,q})^{h_k}$, and

$$S_{s,q} = Q\,\mathrm{diag}(b^2_{s,q,1}, \ldots, b^2_{s,q,4})Q^\top, \tag{29}$$

$$E_{s,aa',q} = (v^\top_{s,aa'} S_{s,q} v_{s,aa'})^{1/2}, \tag{30}$$

$$c_{s,q} = \max_{a<a':\, D_{s,aa'}>0} \frac{D_{s,aa'} - t_q}{E_{s,aa',q}}, \quad G_{s,q} = c^2_{s,q} S_{s,q}, \tag{31}$$

$$\widehat{d}_{s,q}(y, y') = ((y - y')^\top G_{s,q}(y - y'))^{1/2}. \tag{32}$$

This construction attains one-sided population distortion $\epsilon^{\downarrow}_{M,s,q} = q\delta$. The corruption grid is $q \in \{0, 0.2, 0.5, 1, 2\}$ with $\delta = 2$; the primary comparison is $q = 0$ versus $q = 0.5$. Seed-level changes in ARI, PD, and NR use two-sided equivalence margins 0.05, 0.05, and 0.02, respectively. Certificate applicability is evaluated separately from empirical robustness.

Tables 13 and 14, together with Fig. 6, report the primary changes and complete corruption-grid diagnostics.

Table 13: Corruption study: primary seed-level changes for $q = 0$ versus $q = 0.5$ ($n = 20$).

Effect	Mean	Sample SD	95% CI
ΔARI	0	0	[0,0]
ΔPD	0	0	[0,0]
ΔNR	0	0	[0,0]

Each of the three no-change event counts is 20/20, so the corresponding event probability for a positive change has a two-sided Wilson 95% interval $[0, 0.161]$.

The raw utility bound is $B_{s,q} = L(\delta + \epsilon^{\downarrow}_{M,s,q}) = L\delta(1 + q)$ with $L = e^{-1/2}/2$; its non-vacuity is assessed against each task's attainable raw-regret range. At $q = 0$ and 0.2, all 320 applicable

Table 14: Corruption study: complete corruption trajectory. Means use 20 seed-level units. The bound is in raw utility units, unlike dimensionless NR.

q	ARI	FMR	FSR	PD	Blocks	Mean NR	Raw utility bound	Certificate
0	1	0	0	0	8	0	0.6065	non-vacuous; 160/160, 0 violations
0.2	1	0	0	0	8	0	0.7278	non-vacuous; 160/160, 0 violations
0.5	1	0	0	0	8	0	0.9098	vacuous
1	1	0	0	0	8	0	1.2131	vacuous
2	1	0	0	0	8	0	1.8196	vacuous

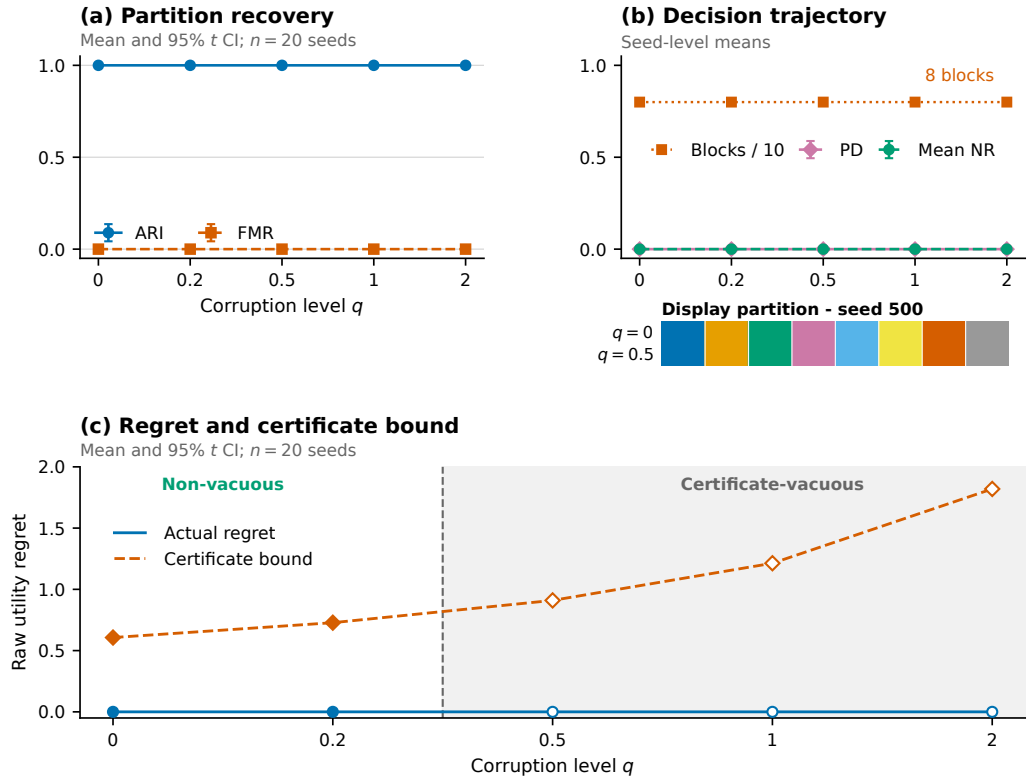


Figure 6: Corruption study: partition, decision, regret, and certificate diagnostics over the complete corruption grid, with $q = 0.5$ as the primary corruption level. Tasks are aggregated within seed. Actual regret and its bound are compared in raw utility units; NR is a separate dimensionless view. At $q = 0$ and 0.2, all 320 applicable certificate checks passed; the certificate is vacuous at $q = 0.5, 1, 2$.

non-vacuous task/seed certificate checks had zero observed violations. Corruption was activated in 20/20 seeds at every $q > 0$.

L Task-Scoped Online DEAQ: Design and Pilot Study

L.1 Runtime interface

At each decision, Online DEAQ constructs features from recent observations, observed ego-motion, and validity information within the pre-action condition e_t, plus the current task. Its 33-dimensional current sector-belief summary h_t contains three 11-bin beliefs in left–centre–right order. The predictor feature $x_t = (h_t, h_t - h_{t-1}, u_t, m_t) \in \mathbb{R}^{72}$ has respective dimensions 33, 33, 4, and 2; it is constructed from e_t without assuming that conditioning on x_t recovers the population law conditional on e_t. Concatenating a seven-way action indicator gives a 79-dimensional input to a $79 \to 128 \to 128 \to 33$ network. Its output softmax normalizes each 11-bin sector separately. Query q_{C} selects the centre belief (sector index 1), and q_{L} selects the left belief (sector index 0). For projected current belief $b_{t,q}$ and predicted post-action belief point $\widehat{b}_{t,a,q}$, the implementation forms the signed update $\widehat{b}_{t,a,q} - b_{t,q}$ and applies half-L_1 distance between two actions' updates, as in Equation (12). The pre-action baseline cancels in the pairwise distance; no uniform baseline of the kind used in the controlled Habitat study is assumed. This operational point distance is not the unknown population-law W_1, and the module uses no global latent state or world model and has no access to the reference partition at runtime. This distance also equals W_1 between Dirac laws at the two predicted update vectors under the half-L_1 ground metric, or W_1 on the 11-bin alphabet with zero–one bin cost. The latter bin-level transport is distinct from population $W_1(P^q_{a|e}, P^q_{a'|e})$ between possibly stochastic update-vector laws.

The online pilot uses the same seven actions as the controlled Habitat study (Appendix J): A00 is stationary RGB/depth acquisition with no simulator step; A01/A02 pan only the sensor node left/right by 10°; A03/A04 translate the body left/right by 0.05 m; A05 translates the body forward by 0.10 m; and A06 yaws the whole body left by 10°. We call these counterfactual simulator branches Oracle branches; they define the reference partition used for evaluation and, on Train-fit episodes, the supervised training targets. For each of 180 decisions, the collector evaluates all seven counterfactual actions from the same restored decision anchor, giving $180 \times 7 = 1{,}260$ Oracle branches. All six Online trajectory and partition records are persisted first. Only afterward are Oracle compact artifacts generated and read for evaluation. One realized branch per action defines the evaluation reference partition.

L.2 Predictor and partition rule

Relation-aware training. One supervised group g contains the seven actions for one episode, decision, and active task. With predicted and Oracle target sector beliefs $\widehat{b}_{g,a,u}$ and $b_{g,a,u}$, the belief term is the mean of the three sector-wise total-variation distances, averaged over all seven actions and training groups:

$$\mathcal{L}_{\mathrm{bel}} = \frac{1}{7 \cdot 3} \sum_{a=0}^{6} \sum_{u=0}^{2} \frac{1}{2} \|\widehat{b}_{g,a,u} - b_{g,a,u}\|_1, \tag{33}$$

with the outer minibatch mean implicit. After the active-task projection, let $\widehat{d}_g(a, a')$ be the differentiable implementation of the signed-consequence half-L_1 distance and let $d_g^{\mathrm{Oracle}}(a, a')$ be the same production distance computed from the Train-fit post-action targets. The relation term and total objective are

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{21} \sum_{0 \le a < a' \le 6} \left| \widehat{d}_g(a, a') - d_g^{\mathrm{Oracle}}(a, a') \right|,$$
$$\mathcal{L} = \mathcal{L}_{\mathrm{bel}} + 1.0\, \mathcal{L}_{\mathrm{rel}}. \tag{34}$$

The relation weight is 1.0. Only the 12 Train-fit episodes estimate normalization and update parameters. The 70 continuous context coordinates use Train-fit population means and standard deviations ($\mathrm{ddof} = 0$) with a 10^{-6} standard-deviation floor; the two validity masks and seven-way action indicator are not standardized. The four disjoint Train-validation episodes are used only for no-gradient epoch evaluation, early stopping, and checkpoint selection. Neither reserved nor evaluation episodes enter normalization, training, stopping, or selection.

All three seeds 8601, 8602, and 8603 use the same $79 \to 128 \to 128 \to 33$ ReLU architecture, Adam with learning rate 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, no scheduler, at most 200 epochs, shuffled Train-fit groups, no dropped final batch, and no worker processes. Integer division of the configured action batch size 256 by seven gives 36 complete decision groups (252 action instances) per update. For each seed, epochs are ordered lexicographically by (Train-validation FMP, Train-validation PPE, epoch, model seed), and training stops after 20 consecutive epochs without improvement under that ordering. The three per-seed best checkpoints are fixed under these common settings, without retraining or selection after the six-episode pilot results. Train-fit Oracle relations are offline supervised targets; evaluation Oracle beliefs and partitions are inaccessible during Online trajectory generation and are revealed only after persistence.

Operational partition and optional representatives. An implementation instead applies $\mathrm{CL}_{\delta,\epsilon_{\mathrm{alg}}}(G)$ to its declared input matrix G: the controlled studies use their respective estimated distances, Static initialization uses the seed-8603 predictor distances, and authorized Online DEAQ refreshes use the cache-dependent $\widetilde{D}_t$ of Equation (13). The operational ensemble contains the three differently seeded frozen relation-aware predictors. The predictors share architecture, Train-fit data, objective, normalization, action set, task projection, and partition threshold.

Static and Online DEAQ use exact copies of the same seed-8603 predictor-derived initialization for each task. Static retains that partition unchanged. Online DEAQ maintains a task-specific cache and permits cache writes only at initialization or an observed task change; non-task-change decisions perform no prediction or repartitioning even when local pre-action information changes. The procedures differ both in task-change refresh and in the three-checkpoint consensus used at refresh; the ablation in Section 5.3 separates the two effects.

For each separated cached pair, the consensus distance is the maximum of the three predictor distances. If that pair ends merged, all three distances must be at most $\theta = \delta + \epsilon_{\mathrm{alg}}$; this pairwise condition is necessary but not sufficient for a complete-linkage merge. For each merged cached pair, the consensus distance is their median. A median above θ is sufficient for that pair to end split, but not necessary: another action can prevent their joint block. One deterministic complete-linkage pass on this cache-dependent dissimilarity then forms the operational partition. The values are $\delta = 0.05$ and $\epsilon_{\mathrm{alg}} = 10^{-12}$, and only the active task cache is replaced. The consensus matrix need not itself be a metric or a population W_1 matrix. The optional representative for a block minimizes the sum of predicted distances to its other members, breaking ties by the smaller canonical action index. The evaluated implementation records only the partition and K; representative selection enters neither PPE, FMP, ARI, compression, nor a downstream task evaluation.

PPE, FMP, and ARI aggregation. A false merge is a pair that the method groups but the reference separates, and a false split is the converse. PPE is the fraction of the $\binom{7}{2} = 21$ unordered action pairs that are false merges or false splits, FMP is the fraction that are false merges, and the false-split proportion equals PPE − FMP. For ARI, let n_{ij} be the contingency table between Oracle and method blocks at one decision, with row and column sums r_i and c_j. Writing $T = \binom{7}{2} = 21$ and $E = (\sum_i \binom{r_i}{2})(\sum_j \binom{c_j}{2})/T$, the implementation computes

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - E}{\frac{1}{2}\left(\sum_i \binom{r_i}{2} + \sum_j \binom{c_j}{2}\right) - E}. \tag{35}$$

For valid seven-action partitions, the denominator is zero only when both partitions are all singletons or both are single blocks; ARI is then defined as 1. Equality here means equal pairwise co-membership and is invariant to block-label permutations. PPE, FMP, and ARI are first computed for each decision. Episode rows average 30 decisions, scene rows average the 60 decisions in that scene's two episodes, and the overall row averages all 180 decisions.

L.3 Pilot suite identity

The six-episode pilot suite contains 180 decisions, 12 task changes, and 1,260 decision branches. It uses two episodes from each of three fixed HM3D scenes. The shorthand mapping is

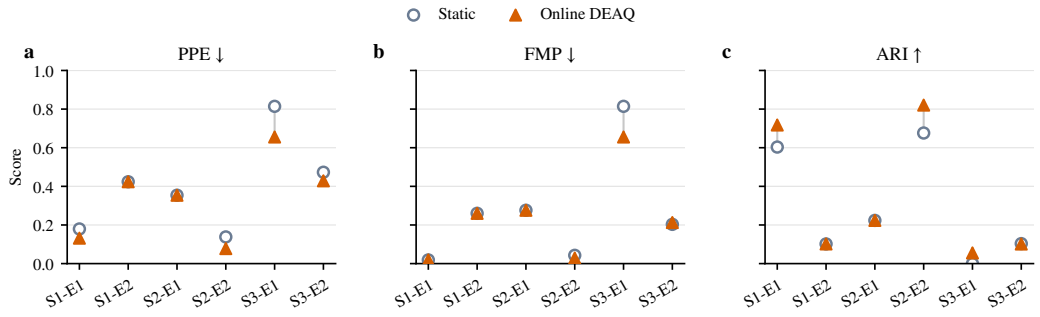


Figure 7: **Per-episode agreement with the reference partition.** Each pair of markers compares Static and Online DEAQ in the same episode; lower PPE and FMP and higher ARI are better. The two methods coincide in S1-E2 and S2-E1, where no refresh changed the partition.

Shorthand	Episode identity
S1-E1	`00802-wcojb4TFT35::p0000`
S1-E2	`00802-wcojb4TFT35::p0001`
S2-E1	`00804-BHXhpBwSMLh::p0000`
S2-E2	`00804-BHXhpBwSMLh::p0001`
S3-E1	`00803-k1cupFYWXJ6::p0000`
S3-E2	`00803-k1cupFYWXJ6::p0001`

L.4 Per-scene and per-episode results

Table 15 reports per-scene means over 60 decisions, Table 16 reports per-episode means over 30 decisions, and Figure 7 plots the per-episode values. The two procedures give identical PPE, FMP, and ARI in S1-E2 and S2-E1 and differ in the other four episodes.

Table 15: Per-scene results for all three scenes in the six-episode pilot suite. FM and FS are total false merges and false splits within a scene.

Scene	Method	PPE	FMP	ARI	FM	FS	Mean K	Compression
00802-wcojb4TFT35	Static	0.301587	0.139683	0.352638	176	204	3.166667	0.547619
00802-wcojb4TFT35	Online DEAQ	0.277778	0.139683	0.409724	176	174	3.000000	0.571429
00804-BHXhpBwSMLh	Static	0.246032	0.159524	0.449904	201	109	2.833333	0.595238
00804-BHXhpBwSMLh	Online DEAQ	0.215873	0.152381	0.522176	192	80	2.833333	0.595238
00803-k1cupFYWXJ6	Static	0.643651	0.508730	0.051768	641	170	1.333333	0.809524
00803-k1cupFYWXJ6	Online DEAQ	0.542063	0.434127	0.078396	547	136	1.500000	0.785714

Table 16: Per-episode fidelity for all six episodes used in the pilot study. No episode is omitted.

Episode	Static PPE	Online PPE	Static FMP	Online FMP	Static ARI	Online ARI
S1-E1	0.179365	0.131746	0.019048	0.019048	0.603446	0.717618
S1-E2	0.423810	0.423810	0.260317	0.260317	0.101829	0.101829
S2-E1	0.353968	0.353968	0.276190	0.276190	0.223642	0.223642
S2-E2	0.138095	0.077778	0.042857	0.028571	0.676165	0.820711
S3-E1	0.814286	0.655556	0.814286	0.655556	0.000000	0.055377
S3-E2	0.473016	0.428571	0.203175	0.212698	0.103535	0.101414

L.5 REFRESH DIAGNOSTICS

Across 12 task changes, the operational partition changed four times and remained unchanged eight times. Updates produced 11 newly merged pairs and 12 newly split pairs; 57 action pairs exhibited disagreement among the three predictors. Predictor inference counts were 12 for seed 8601, 12 for seed 8602, and 18 for seed 8603, for 42 total inferences. In the recorded diagnostic, Static initialization took 0.020359 s and Online ensemble maintenance took 0.032177 s.