SUPERVISION BEFORE STRUCTURE: WHAT A WORLD MODEL NEEDS TO KNOW WHAT IT CANNOT ANSWER

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model asked a counterfactual question should first decide whether its evidence determines the answer at all. We study what such a model needs for that decision on a benchmark where the answer is known exactly: ten environments over seven systems (six physical, one structural causal model) whose hidden mechanism is drawn from a finite bank, so compatible laws, answer dispersion and the value of every probe are enumerated rather than estimated. We test the causal world-model design the field is converging on — a mechanism posterior with per-mechanism counterfactual heads — against a supervision-matched control with none of that machinery. Three findings follow. With an answerability label on every training query, learned structure is redundant: eight of nine ablations are null, the posterior's own dispersion scores answerability far below a learned head, and the one margin the architecture showed was an artifact of the baseline's training loss. With scarce labels — as few as forty — a factorised posterior improves label efficiency by up to 0.06 AUROC, needs no mechanism labels, and transfers to a non-physics causal model; auxiliary supervision, generic factorised capacity, independent heads and width do not account for it, nor, on most environments, do the posterior's own auxiliary losses. Acting on the query is at oracle level for both models; stopping and answering after self-gathered evidence are weak for both. We release the benchmark, every control, and the failed hypotheses.

1 INTRODUCTION

Consider an agent that watches a ball strike a switch and sees the switch open. Six physical laws are consistent with what it saw. *What if the ball had hit harder?* — every consistent law agrees: the switch opens. *What if it had hit softer?* — the laws disagree completely. Same evidence, same episode; one question is answerable and the other is not (Figure 1). A model that answers the second confidently has answered a question the evidence does not determine, and no accuracy metric on answerable questions will catch it.

Recent work has argued that world models should carry explicit causal structure precisely so that they can recognise such situations: a posterior over mechanisms whose disagreement signals that a counterfactual is undetermined, and an evidence policy that resolves the disagreement relevant to the query (Zhou & Zhang, 2026; Tan et al., 2026; Go et al., 2026; Fotias, 2026). We built it, and this paper reports what happened when we tested it against the control that the closest prior work explicitly defers (Zhou & Zhang, 2026): a plain encoder given the same identifiability labels and none of the causal machinery. Our object is this architecture family — a categorical posterior over a finite bank read by per-mechanism heads; continuous latent-parameter designs (Tan et al., 2026; Zhang et al., 2026) are not tested. Throughout, *identifiability supervision* means the direct query-level label "is this counterfactual determined by the evidence"; both models also receive counterfactual-answer and evidence-gain targets, the structured model additionally the mechanism-family target, and when we withhold labels only the first is withheld.

Testing the design requires ground truth that ordinary simulators cannot supply, so we build our study on a benchmark, ID-Phys++, in which every identifiability label is exact by construction (Section 2): the space of mechanisms is a finite bank, so compatibility, answer dispersion and the value of every probe are computed by enumeration. That lets the comparison be made cleanly and repeatedly — which turned out to matter, because the first answer we got was wrong.

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.

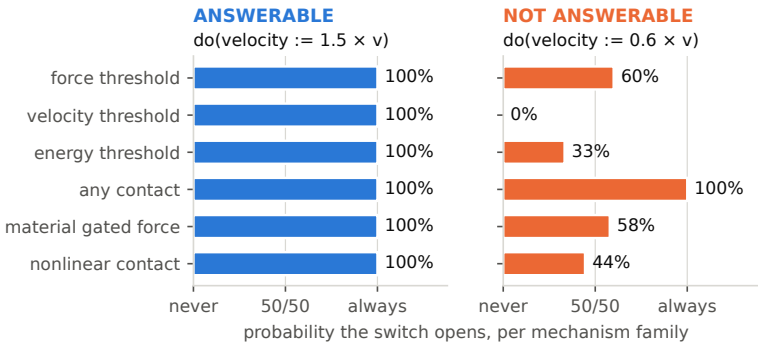

Figure 1: Identifiability is a property of the question. One observed episode; all six mechanism families remain compatible. Each bar is one family's counterfactual answer. Strengthening the cause is answerable because every surviving law agrees; weakening it is not. Generated from the simulator.

Contributions.

1. **Under full supervision, learned structure is redundant.** With a supervision-matched control and a direct identifiability label on every training query, none of the tested learned components improves identifiability prediction: eight of nine ablations are null, the posterior's own dispersion scores answerability far below a learned head, and an exact posterior in place of the learned one adds headroom only on the ramp system (Section 3).
2. **A structural margin that was a training artifact.** The architecture's one-point lead on two environments, robust to pixel noise, disappears once both models are trained with a tolerance-scaled counterfactual loss; a comparison against a baseline is a comparison against how both were trained (Section 4).
3. **Acting on the query: the same control, the same answer.** As a sequential agent the model selects probes within a few percent of the exact oracle, and so does the control once its evidence head is trained on the same target; stopping and answering after probing are weak for both and improve with the agent's own probe histories (Section 5).
4. **Where structure pays: scarce identifiability supervision.** With 1% of the labels the explicit posterior leads the control on every environment ($+0.003$ to $+0.021$ AUROC), with 0.1% on six of eight (up to $+0.036$, and up to $+0.065$ for the variant without mechanism labels), and at full supervision by nothing — both models at their final checkpoint, no validation label used (Table 1). Six further controls fail to explain the margin: auxiliary mechanism supervision, a generic factorised head, its gate statistics, independent heads, twice the width, and the posterior's own dispersion and calibration losses do not account for it, while removing the mechanism supervision or randomising its categories never reduces it and at 0.1% often helps. The cause is the per-mechanism reduction trained as a posterior; we could not isolate it further. The variant without mechanism labels is the one we recommend (Section 3).
5. **The benchmark and every control**, including those that overturned our own claims (Section 2, appendix).

2 Setup: Exact Identifiability Labels

Environments and the mechanism bank. Ten environments over seven systems (seven state-observed, three rendered at $6 \times 32 \times 32$): a ball striking a switch, a ball rolling off a ramp, a block on a spring, a pendulum whose string catches a peg, a striker launching a second ball toward a cup, rendered versions of the first three (so that two systems exist as state/pixel pairs), and a spring whose dissipation is one *continuous* four-parameter composite law with a bank of $3{,}200$ sampled parameter points rather than a grid of functional forms (its four "families" are the regions of parameter space in which one term dominates, so the $K{=}4$-slot posterior is a posterior over regions); and one non-

physics domain, a dose-response *structural causal model*: a latent mediator $Z := f_Z(\text{dose}, \text{baseline})$ drives the observed response $Y = \text{gain} \cdot Z$, the hidden mechanism is f_Z (linear, saturating, Hill, threshold, multiplicative, biphasic), and a clamp $\text{do}(Z{:=}c)$ cuts the ambiguous edge. Each carries a finite *mechanism bank* $\mathcal{M}$: six causal families instantiated on a parameter grid finer than the measurement tolerance. Finite banks are the design decision that buys exactness: everything below is computed by enumeration.

Labels. For an initial condition c (a *control* in the simulator's sense) and observed outcome o, the compatible set is $\mathcal{C} = \{m \in \mathcal{M} : \text{sim}(m, c) \simeq o\}$. For a query $q = (\text{do}(X{:=}x'), \text{target})$ we evaluate every $m \in \mathcal{C}$, cluster the answers under complete linkage with diameter bounded by the target's tolerance, and define D_q as their dispersion (for binary targets the Gini impurity of the answers; for scalar targets the mean pairwise absolute difference in units of the tolerance) and $\textsf{identifiable}(q)$ as "exactly one answer class". For every candidate probe we enumerate its possible outcomes and record its exact identifiability gain, a generic information gain, and a resolution-aware structural gain in the sense of Fotias (2026). Episodes are retained only if genuinely aliased (≥ 2 families, ≥ 4 hypotheses compatible), and labels that rest on a single grid point are rejected. The benchmark has $\sim$330K queries, 30–47% non-identifiable per environment, with IID, parameter-, mechanism-, composition- and (for rendered environments) visual-OOD splits, leakage-checked programmatically (Appendix A).

Reference architecture (QID-AWM). A transformer over the observation history, FiLM-conditioned on the query; a categorical posterior over the bank; a per-mechanism counterfactual head mixed under that posterior; an identifiability head fed the belief, the query and the posterior's own dispersion statistics; and an evidence head trained to rank probes by exact oracle gain — the design the literature proposes, and the one we set out to validate.

Controls. B1–B4 are a deterministic world model, a five-member deep ensemble, MC dropout, and ensemble disagreement, none of which see the identifiability label. **B1+ID** is the supervision-matched control: B1's encoder plus a supervised identifiability head — the same labels as QID-AWM, none of its causal machinery. B0 is a query-type prior that never sees the observation. Unless a table states otherwise, results are means over five seeds; the plain control at 0.1% and 1% and the visual-ramp extrapolation use ten, and a few auxiliary controls use three or eight, always shown as superscripts. Differences are reported as paired effect sizes with 95% bootstrap intervals over seeds, and two models are called *equivalent* ($\equiv$) when the whole interval lies within ± 0.002 AUROC, *different* ($\neq$) when it excludes zero and is not equivalent, and uncertain ($\sim$) otherwise (five-seed intervals describe seed-to-seed variation only). With unequal seed counts, differences and intervals use the shared seeds; cell means use all.

3 Supervision before structure

Appendix Table 17 reports the headline comparison. On the two saturated collision environments every method exceeds 0.999 and nothing can be learned. On `spring_block` the control is equivalent to the full model (within 0.001 AUROC on every split). On the two ramp environments the full model leads by $+0.009$ AUROC (five seeds, every seed between $+0.007$ and $+0.012$) — we return to this in Section 4. The null is not an IID artifact: on the parameter-, mechanism- and composition-OOD splits (Appendix Table 18) the two models move together, with one exception worth stating as a positive: parameter extrapolation from pixels (`visual_ramp_ball`), where over ten seeds the structured model scores 0.923 ± 0.005 against the control's 0.898 ± 0.025 ($+0.025$, $[+0.011, +0.041]$) while the two are equivalent IID: factorisation does not help this system in distribution, but it extrapolates better, and five times less variably. Pixel-appearance shift costs both models 0.005–0.07 AUROC but 0.34–0.38 of counterfactual accuracy on the visual ramp (tolerance-scaled loss, five seeds; Appendix Table 19): the answer head, not the identifiability head, is appearance-sensitive. Nor is the comparison a query-type shortcut: the type prior B0 reaches 0.74–0.90 on extreme base rates, but on the balanced types (16–68% of test queries) it falls to 0.53–0.68 while both models keep 0.91–1.00 and stay equivalent (Appendix Table 9); the SCM environment has no balanced type (clamp queries are always answerable, scalar dose queries almost never), so B0 reaches 0.95 there and that domain says little about query difficulty — both learned models still exceed the shortcut at 0.995.

Under full supervision, no learned component is load-bearing. Appendix Figure 4 shows nine ablations of the architecture. Eight move identifiability AUROC by ≈ 0.000 (mean -0.002 to $+0.001$ across environments). Exactly one matters: removing identifiability supervision costs 0.405 (raw loss; 0.35–0.58 under the tolerance loss, Appendix Table 25). Withholding the query entirely costs 0.32–0.41 (0.27–0.34 under the tolerance loss): the *query* is essential, early *conditioning* on it is not. The pattern holds under the tolerance-scaled loss on the two systems added for this paper (Appendix Table 25), and the null holds at d=128/2 layers, d=512/6 layers and d=1024/6 layers, at double the training length, and with 2 to 24 posterior particles: the gap stays within ± 0.0016 AUROC (Appendix Table 27).

The posterior's dispersion is not the signal. The design premise is that answerability can be read off the mechanism posterior as $D_q < \epsilon$. Scoring each query by the model's predicted dispersion across mechanism hypotheses — no learned head, rank-normalised per target type, which favours it — gives AUROC 0.690 on `ramp_ball`, 0.661 on `spring_block` and 0.693 on `visual_ramp_ball`, against 0.978–0.984 for the learned head (Appendix Table 29). Part of the reason is the posterior itself: scored against the exact compatible set (Appendix Table 5), it places 0.75–0.99 of its mass on compatible families against 0.62–0.94 for the family prior, ranks a compatible family first 82–100% of the time, and its predicted dispersion rank-correlates with the exact one at only 0.38–0.83. The architecture, trained by these losses, does not by itself produce a posterior sharp enough to read answerability from; the identifiability signal lives in the supervised head.

What a perfect posterior would add. Replacing the learned posterior by the exact compatible family set from the bank — uniform over compatible families, not grid points (one exact family-level reference, not the unique posterior the bank induces); read by the answer reduction, the dispersion statistics and the identifiability head alike, the family loss switched off — bounds what better posterior training could buy (Appendix Table 10). Across the eight non-saturated environments, at full supervision the bound is $+0.014$–0.015 on the two ramp environments, $+0.003$–0.004 on the three spring environments, and nothing on the pendulum, two-body and causal-model environments, which sit at 0.992–0.997 already. The null therefore reads differently by environment: on the ramp system the learned posterior leaves real headroom, elsewhere the task leaves none. At scarce labels — both models at their final checkpoint, as throughout the next paragraphs — the exact posterior helps at 0.1% on four of six environments ($+0.011$ on `ramp_ball`, $+0.026$ on the pendulum, $+0.019$ on the continuous spring, $+0.005$ on the two-body system; within ± 0.008 on the spring and the causal model) and nowhere at 1% (within ± 0.007). Posterior accuracy therefore matters only in the scarcest regime, and even there it is not what the structured model's margin consists of: the margin reported next is largest where the exact posterior adds least.

Posterior structure is an inductive bias, visible only when identifiability labels are scarce. The null above is measured with an identifiability label on every training query. Withholding labels from a fixed random subset of queries — the same subset for both models, with every loss that consumes the identifiability label or its derived dispersion (computed by the same enumeration, and masked with it) masked identically, and the counterfactual, mechanism and evidence losses untouched — separates them (Table 1, columns "plain" and "QID-AWM"; Figure 2, left). With 1% of the labels the explicit posterior leads by $+0.003$ to $+0.021$ AUROC on every environment; with 0.1% — about forty labelled queries — by $+0.009$ to $+0.036$ on six of eight (ramp, visual ramp, pendulum, two-body, the continuous spring and the causal model) and by nothing on the two spring environments; with 0.5% by $+0.004$ to $+0.025$ on all eight, the two spring environments included; with 5% by $+0.003$ to $+0.010$; with 10% and 25% by thousandths; and at full supervision by nothing except on `visual_spring_block` (Section 4; Appendix Table 20). With no labels at all both sit at chance. Checkpoint selection is a confound in this regime and we remove it: at every fraction below 100% both models are trained for the same 60 epochs and evaluated at their final checkpoint, so neither sees a validation label. Selecting the structured model by validation identifiability AUROC instead (the validation split is fully labelled) inflates its 0.1% margin by 0.01–0.03 and turns two null cells into apparent $+0.02$ margins, while leaving 1% and 5% unchanged (Appendix Table 7); selection by validation counterfactual accuracy lies in between. Plain cells at 0.1% and 1% have ten seeds on four environments and five on the rest (Table 1).

The source of the gain: never mechanism semantics, sometimes substitutable. The structured model is trained on one signal the control is not: the mechanism-family target. Three further controls ask what, in the structured model, explains the margin (Table 1). *B1+ID+MechAux* gives the plain control the same family target, as an auxiliary head on its encoder with the same loss and weight, so that supervision is matched exactly and only the explicit posterior differs; it recovers little or none of the margin on six of eight environments, half on the two-body and a quarter on the SCM at 1%, and at 0.1% sits below plain on three (by 0.014–0.017) and level on the rest. *B1+KHead* gives the plain control a generic factorised answer head — K expert embeddings, one shared counterfactual head applied to each (the structured model's per-mechanism heads are built the same way, so the two differ in the posterior and the family loss, not in parameters per branch), a softmax gate on the belief, and gate statistics that never reach the identifiability head; for K from 2 to 24 it matches the plain control (Appendix Table 6) and stays below the structured model on every environment. Routing the gate's four disagreement statistics to the identifiability head, as the structured model routes its posterior's (*KHead+gate*), changes nothing: within 0.01 of KHead everywhere, below the structured model on all eight at 1% and four of eight at 0.1%. Nor is it parameter isolation: one independent counterfactual head per expert (*KHead-ind*) leaves KHead below the structured model on thirteen of sixteen cells, by 0.003–0.039, uncertain on two, and above only where the structured model itself trails plain (Appendix Table 11). Nor is width: a plain control at d=512 is no better than the standard one and trails the structured model by $+0.004$ to $+0.023$ at both fractions (three of six 0.1% cells uncertain; Appendix Table 12). *No-mech* drops the family loss from the structured model; it matches the full model at 1% and 5% (within 0.005 except the causal model, -0.009 at 1%) and *exceeds* it at 0.1% on the pendulum ($+0.029$), the continuous spring ($+0.015$), the ramp ($+0.012$) and the visual ramp ($+0.007$): at forty labels the family loss costs accuracy. Over the plain control it leads by $+0.017$ to $+0.065$ on six environments at 0.1% (nothing on the two spring environments) and by $+0.004$ to $+0.012$ on all eight at 1% (Appendix Table 13). Needing no mechanism labels at all, it is the configuration we recommend. *Scrambled* keeps the family loss but trains it on a random category drawn independently per episode, destroying the correspondence between posterior components and physical mechanisms; it, too, matches the full model (within 0.014 at 1% and 5%, five seeds), and where no-mech exceeds the full model by more than 0.005 (four cells at 0.1%) so does it. Two further controls test the remaining candidate directly. Removing the dispersion and calibration terms from the structured model — the two losses that shape its posterior — changes nothing at 1% (within 0.003; the causal model -0.007) and at 0.1% is uncertain everywhere except the pendulum, where removing them helps ($+0.011$); adding them to KHead lifts it at 0.1% on the ramp ($+0.012$) and the causal model ($+0.010$) and nowhere else (Appendix Table 8). Across all ten environments, then, seven candidate ingredients fail: extra supervision does not account for the gain, generic factorised capacity does not, its gate statistics do not, parameter isolation across heads does not, width does not, the posterior-shaping losses do not on four of the six environments tested, and removing the family loss or destroying the categories' meaning does not reduce it. What survives every elimination is the per-mechanism counterfactual reduction trained as a K-component posterior over the history; we could not isolate the ingredient further, and say so. The causal model (last row of Table 1) repeats the physical pattern: at 0.1% no-mech and scrambled (0.940, 0.937) sit with the full model (0.936), far above plain (0.912), while +MechAux (0.898) and KHead (0.910) sit with plain; at 1% likewise, +MechAux recovering a quarter of the $+0.021$ margin. The cells that do not separate are the two spring environments at 0.1%, where no variant leads the plain control. The margin is not specific to identifiability labels either: subsampling identifiability, counterfactual and evidence targets jointly leaves $+0.010$ on the spring at 5% and thousandths elsewhere (Appendix Table 16).

4 A STRUCTURAL MARGIN, AND THE TRAINING LOSS THAT PRODUCED IT

The ramp environments were the exception to the null, and they came with a story. The obvious objection to Section 3 is that our labels are a deterministic function of the clean observation, so answerability prediction is noiseless function approximation and a supervised head needs no posterior. We tested it by adding a stochastic observation channel and re-training both models under identical noise. On state observations the margin vanished at the first noise level. From pixels (`visual_ramp_ball`) it *persisted* at every level from 0.15 to 1.0 — $+0.009$ decaying to $+0.0015$, the bootstrap interval excluding zero at every level, five seeds (Table 3, "raw"). This was, briefly, the paper's headline.

Table 1: What causes the scarce-supervision advantage? (tolerance-scaled loss, IID AUROC, seeds as superscripts; both models at their final checkpoint). *plain*: B1+ID; *+MechAux*: B1+ID given the structured model's mechanism-family target as an auxiliary head; *no-mech*: QID-AWM without that target; *scrambled*: QID-AWM with the target replaced by a random category per episode; *KHead*: B1+ID with a generic K-expert answer head whose gate never reaches the identifiability head; *QID-AWM*: full. Δ is QID-AWM minus plain with its paired-bootstrap verdict ($\neq$: CI excludes zero). Full sweep (0.1–100%) in Appendix Table 20.

	0.1% of identifiability labels							1% of identifiability labels						
env	plain	+MechAux	no-mech	scrambled	KHead	QID-AWM	Δ	plain	+MechAux	no-mech	scrambled	KHead	QID-AWM	Δ
ramp	$0.886^{(10)}$	$0.872^{(5)}$	$0.909^{(5)}$	$0.907^{(5)}$	$0.889^{(8)}$	$0.897^{(5)}$	+0.010 $\neq$	$0.946^{(10)}$	$0.946^{(5)}$	$0.954^{(5)}$	$0.952^{(5)}$	$0.950^{(8)}$	$0.955^{(5)}$	+0.009 $\neq$
spring	$0.864^{(10)}$	$0.870^{(5)}$	$0.859^{(5)}$	$0.858^{(5)}$	$0.857^{(8)}$	$0.861^{(5)}$	-0.004 $\sim$	$0.927^{(10)}$	$0.930^{(5)}$	$0.937^{(5)}$	$0.937^{(5)}$	$0.932^{(8)}$	$0.940^{(5)}$	+0.012 $\neq$
vis. ramp	$0.894^{(5)}$	$0.893^{(5)}$	$0.919^{(5)}$	$0.921^{(5)}$	$0.898^{(5)}$	$0.912^{(5)}$	+0.018 $\neq$	$0.936^{(5)}$	$0.936^{(5)}$	$0.946^{(5)}$	$0.946^{(5)}$	$0.938^{(5)}$	$0.948^{(5)}$	+0.012 $\neq$
vis. spring	$0.855^{(5)}$	$0.854^{(5)}$	$0.857^{(5)}$	$0.843^{(5)}$	$0.848^{(5)}$	$0.850^{(5)}$	-0.005 $\sim$	$0.926^{(5)}$	$0.929^{(5)}$	$0.937^{(5)}$	$0.937^{(5)}$	$0.926^{(5)}$	$0.938^{(5)}$	+0.013 $\neq$
pendulum	$0.818^{(10)}$	$0.830^{(5)}$	$0.886^{(5)}$	$0.877^{(5)}$	$0.823^{(8)}$	$0.856^{(5)}$	+0.036 $\neq$	$0.961^{(10)}$	$0.961^{(5)}$	$0.966^{(5)}$	$0.965^{(5)}$	$0.961^{(8)}$	$0.965^{(5)}$	+0.003 $\neq$
two-body	$0.931^{(10)}$	$0.933^{(5)}$	$0.948^{(5)}$	$0.943^{(5)}$	$0.936^{(8)}$	$0.944^{(5)}$	+0.013 $\neq$	$0.963^{(10)}$	$0.968^{(5)}$	$0.971^{(5)}$	$0.971^{(5)}$	$0.964^{(8)}$	$0.974^{(5)}$	+0.011 $\neq$
spring (cont.)	$0.902^{(5)}$	$0.885^{(5)}$	$0.926^{(5)}$	$0.930^{(5)}$	$0.903^{(5)}$	$0.911^{(5)}$	+0.009 $\neq$	$0.956^{(5)}$	$0.952^{(5)}$	$0.964^{(5)}$	$0.964^{(5)}$	$0.956^{(5)}$	$0.963^{(5)}$	+0.007 $\neq$
SCM dose	$0.912^{(5)}$	$0.898^{(5)}$	$0.940^{(5)}$	$0.937^{(5)}$	$0.910^{(5)}$	$0.936^{(5)}$	+0.024 $\neq$	$0.951^{(5)}$	$0.957^{(5)}$	$0.963^{(5)}$	$0.964^{(5)}$	$0.952^{(5)}$	$0.973^{(5)}$	+0.021 $\neq$

The control that overturned it. Every model had been trained with a Huber loss on the scalar counterfactual in raw physical units, which weighs a 0.2 m error on a 20 m roll the same as on a 0.2 m one. That is a training choice, not a property of either model. We re-trained both models with each scalar error expressed in units of that query's answer tolerance — the quantity that defines a correct answer — on every environment, five seeds (Table 2). On the five original environments the margin disappears: -0.0001 on `ramp_ball` and $+0.0006$ on `visual_ramp_ball`, both inside the equivalence region. The two environments lose it in opposite ways: on `ramp_ball` the full model *falls* to the control's level ($0.993 \rightarrow 0.981$), so the raw loss had been flattering the architecture; on `visual_ramp_ball` the control *rises* ($0.968 \rightarrow 0.977$), so it had been handicapped. The same loss removes the structured model's counterfactual-accuracy deficit under the raw loss (0.74 vs 0.82 on the causal model, 0.87 vs 0.91 on the two-body system; Appendix Table 33), leaving the two models within 0.02 everywhere (Table 2). Optimiser settings are identical for both models and both losses, and the equivalence holds at a third and at three times the learning rate (Appendix Table 23).

Repeating the noise sweep under the tolerance-scaled loss completes the picture, and complicates it (Table 3, "tol."; Figure 2, right). On both ramp environments the margin never returns: between -0.0006 and $+0.0008$ at every noise level. On `spring_block`, which had shown no margin under the raw loss at any noise level, a margin now *appears* under noise and grows with it: $+0.003$ at $\sigma = 0.15$ to $+0.009$ at $\sigma = 1.0$, five seeds, the bootstrap interval excluding zero at every level. Both models degrade with noise, the control faster (Appendix Table 14).

A second state/pixel pair. That result suggested a test: render `spring_block` to pixels, as `ramp_ball` already was. Under the tolerance-scaled loss, `visual_spring_block` shows the margin on *clean* observations, $+0.005$ ($[+0.003, +0.007]$, five seeds; $+0.010$ under the raw loss). Adding pixel noise then removes it: $+0.004$ at $\sigma = 0.15$, $+0.003$ at 0.30, and ≤ 0.001 at 0.50 and 1.0 (five seeds). One reading fits both channels: where the observation channel is hard but not hopeless (noisy state, clean pixels) an explicit posterior helps by a few thousandths; clean state and degraded pixels leave nothing for it to add. On the ramp system the same sequence yields nothing, and the collision system is saturated: under the recipe we recommend, the explicit posterior earns a small margin on one of six physical systems.

Four recipes, not two. Two more scalar recipes make the point a matrix rather than a pair (Appendix Table 22): error relative to the answer's magnitude, and a Huber on a symmetric log transform of both sides. On `spring_block`, the pendulum and the two-body system every recipe leaves the two models equivalent (95% CI inside ± 0.002); on `visual_spring_block` every recipe leaves a margin ($+0.005$ to $+0.010$). On the two ramp environments the raw loss gives $+0.009$, the tolerance loss ± 0.001, and the relative and log recipes $+0.005$ to $+0.010$: three of four recipes show a margin, and the one whose units are the benchmark's own definition of a correct answer shows none, which is why we report it as the principled choice.

We draw the general lesson rather than the local one: whether an architectural component "helps" is a statement about a difference between two trained models, and the training recipe is part of that difference. Changing one loss term moved the margin off the two environments that had shown it,

Table 2: Supervision-matched comparison under both training losses (IID; Δ is QID-AWM minus B1+ID with seeds as superscript and its paired-bootstrap verdict: $\equiv$ interval within ± 0.002, $\neq$ interval excludes zero, $\sim$ otherwise). The apparent structural margin on the ramp environments belongs to the loss.

	raw-unit scalar loss (reported)			tolerance-scaled scalar loss			
env	QID-AWM	B1+ID	Δ (verdict)	QID-AWM	B1+ID	Δ (verdict)	cf-acc Q/B
collision	0.9997	0.9996	$+0.0001^{(5)} \equiv$	0.9997	0.9996	$+0.0001^{(5)} \equiv$	0.938/0.938
ramp	0.9931	0.9837	$+0.0094^{(5)} \neq$	0.9806	0.9807	$-0.0001^{(5)} \equiv$	0.841/0.843
spring	0.9828	0.9829	$-0.0001^{(5)} \equiv$	0.9836	0.9840	$-0.0004^{(5)} \equiv$	0.922/0.923
vis. collision	0.9997	0.9996	$+0.0001^{(5)} \equiv$	0.9996	0.9996	$+0.0000^{(5)} \equiv$	0.939/0.938
vis. ramp	0.9772	0.9684	$+0.0088^{(5)} \neq$	0.9776	0.9768	$+0.0008^{(10)} \equiv$	0.808/0.822
vis. spring	0.9764	0.9667	$+0.0097^{(5)} \neq$	0.9772	0.9721	$+0.0051^{(5)} \neq$	0.913/0.913
pendulum	0.9927	0.9929	$-0.0002^{(5)} \equiv$	0.9922	0.9924	$-0.0003^{(5)} \equiv$	0.899/0.897
two-body	0.9966	0.9968	$-0.0001^{(5)} \equiv$	0.9970	0.9970	$+0.0000^{(5)} \equiv$	0.967/0.971
spring (cont.)	0.9876	0.9866	$+0.0010^{(5)} \equiv$	0.9878	0.9876	$+0.0002^{(5)} \equiv$	0.933/0.934
SCM dose	0.9956	0.9956	$+0.0000^{(5)} \equiv$	0.9949	0.9943	$+0.0005^{(5)} \equiv$	0.877/0.892

Table 3: QID-AWM minus B1+ID (identifiability AUROC) across observation-noise levels, under the reported raw-unit loss and the tolerance-scaled loss. Superscripts give seeds.

env	loss	noise 0	noise 0.15	noise 0.3	noise 0.5	noise 1
ramp	raw	$+0.0109^{(5)}$	$+0.0004^{(5)}$	$-0.0001^{(5)}$	$-0.0000^{(5)}$	$-0.0001^{(5)}$
ramp	tol.	$-0.0001^{(5)}$	$-0.0001^{(5)}$	$-0.0000^{(5)}$	$-0.0001^{(5)}$	$-0.0000^{(5)}$
vis. ramp	raw	$+0.0090^{(5)}$	$+0.0083^{(5)}$	$+0.0065^{(5)}$	$+0.0049^{(5)}$	$+0.0015^{(5)}$
vis. ramp	tol.	$+0.0008^{(10)}$	$+0.0000^{(5)}$	$-0.0004^{(5)}$	$-0.0006^{(5)}$	$-0.0006^{(5)}$
spring	raw	$+0.0002^{(5)}$	$+0.0003^{(5)}$	$+0.0003^{(5)}$	$-0.0002^{(5)}$	$+0.0000^{(5)}$
spring	tol.	$-0.0004^{(5)}$	$+0.0030^{(5)}$	$+0.0048^{(5)}$	$+0.0070^{(5)}$	$+0.0094^{(5)}$
vis. spring	raw	$+0.0097^{(5)}$	$+0.0124^{(5)}$	$+0.0118^{(5)}$	$+0.0079^{(5)}$	$+0.0020^{(5)}$
vis. spring	tol.	$+0.0051^{(5)}$	$+0.0043^{(5)}$	$+0.0031^{(5)}$	$+0.0008^{(5)}$	$+0.0003^{(5)}$
pendulum	raw	$-0.0002^{(5)}$	$-0.0007^{(5)}$	$-0.0003^{(5)}$	$-0.0005^{(5)}$	$-0.0001^{(5)}$
pendulum	tol.	$-0.0003^{(5)}$	$-0.0002^{(5)}$	$-0.0000^{(5)}$	$+0.0001^{(5)}$	$-0.0003^{(5)}$
two-body	raw	$-0.0001^{(5)}$	$-0.0003^{(5)}$	$-0.0006^{(5)}$	$-0.0010^{(5)}$	$-0.0002^{(5)}$
two-body	tol.	$+0.0000^{(5)}$	$-0.0001^{(5)}$	$-0.0000^{(5)}$	$-0.0003^{(5)}$	$-0.0003^{(5)}$

onto one that had not, and changed its sign of dependence on noise; two further recipes put a margin back on one of them. The tolerance-scaled loss is the one we recommend for any baseline compared on this benchmark. It leaves two defensible advantages, neither large: at most 0.009 in distribution on the spring system when its observations are hard, and the $+0.025$ parameter-extrapolation margin of the visual ramp system (Section 3).

5 Acting on the query: the same control, the same answer

If the causal machinery does not help the model *know* whether a query is answerable, does it help the model *do* something about it? The benchmark's exact probe values make this measurable as regret against the true optimum. We run the trained model as a sequential agent: at each step it re-encodes its history (factual trial plus every probe so far), ranks the menu with its evidence head, executes the chosen probe in the true world, and repeats until the query is resolved or a four-probe budget is spent. Four oracle policies run on identical menus: exact query-specific gain, generic information gain, RAED-style structural gain, and random. The control runs the identical rollout: B1+ID carries the same evidence head, trained on the same oracle-gain ranking target, with no mechanism posterior behind it — the action-side analogue of the supervision-matched comparison of Section 3.

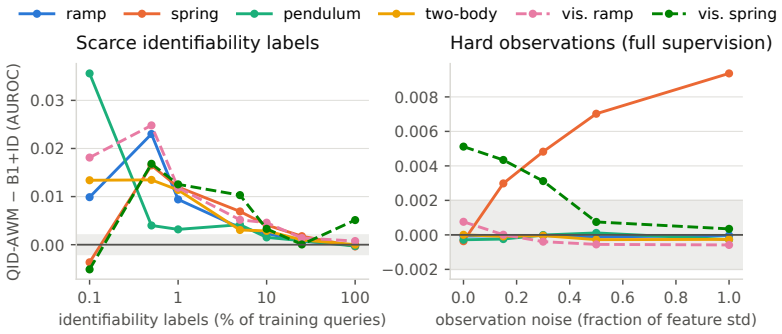

Figure 2: Where posterior structure pays. QID-AWM minus B1+ID identifiability AUROC under the tolerance-scaled loss, against the fraction of identifiability labels (left; 100% is the main run) and against observation noise at full supervision (right). Solid: state observations; dashed: pixels; grey band: the ± 0.002 equivalence region. Structure pays when direct supervision is scarce; at full supervision, in distribution, only on the spring system and only when its observation channel is hard.

Action selection is at oracle level — for both models. Appendix Table 4: on all ten environments the learned agent needs 1–10% more probes than the exact query-specific oracle and, where the oracle's own margin is not marginal, recovers most of the oracle's saving over generic and structural information gain — 40% on the saturated collision environments, 23% on the two-body system (where the oracle saves 30%), 7–17% elsewhere, the interval on the probe difference excluding zero on every one of them (on the causal model the oracle saves only 8% and the agent's 2% is within noise). The control does the same: regret 0.4–8.6% against the full model's 1.0–9.9%, and 8–41% over generic information gain against 7–41%. The pendulum is the exception for both: the oracle itself saves only 2% there, and at 5–6% regret both agents end up 3–5% *worse* than generic. The target, not the posterior, makes the policy query-specific.

Stopping and answering are the unsolved part, for both. Left to its own stopping rule (stop when $p(\text{identifiable}) \geq 0.5$), either agent over-probes — 3.1–3.6 probes where 1.1 suffice — and its final answer is poor even when the compatible set's majority answer is 93–96% correct; the control stops slightly earlier and resolves slightly less, and answers no better (Appendix Table 4). Raising the threshold trades probes for resolution but never moves answer accuracy (Appendix Table 26), so the weakness is representational: no training history ever contained a probe, and both the identifiability head and the answer head are off-distribution after one; the same gap in exploration and belief updating separates humans from frontier models on interactive world-model queries (Warrier et al., 2026). Training on histories with executed probes appended repairs part of it. With at most one random probe per history the agent on `spring_block` stops after 2.5 probes instead of 3.5 and answers 0.80 correctly instead of 0.78; on `visual_ramp_ball` it keeps the original calibration while cutting probes from 3.6 to 2.7 and raising answer accuracy from 0.47 to 0.54, and on the pendulum and two-body systems from 0.64 to 0.73 and 0.72 to 0.88 (on the two-body system a single appended probe first lowers answer accuracy to 0.59 and two restore it — an instability we report rather than explain; Appendix Figure 5 plots every variant). Random probes are the wrong histories, so we let the trained model choose the probes in its training histories (on-policy, labels still exact from the bank), which keeps the probe saving and restores the resolution: at threshold 0.9 the on-policy agent resolves more queries than the original with no more probes and answers better on three of four state systems and matches it on the two-body system at 0.86 correct (Appendix Table 26); calibration on ordinary histories is unchanged. What remains is the gap to the oracle-stopped agent's answers, not its stopping.

Calibration, and a failure only exact labels can expose. Where the supervised models do separate from the confidence-based baselines is calibration: ECE 0.004–0.06 against 0.20–0.36 for B1–B3 (tolerance-scaled loss, five seeds; Appendix Table 30). Split-conformal intervals guarantee only marginal coverage, and the exact labels show what that leaves out: IID coverage is 0.88–0.91 overall but 0.77–0.87 on the non-identifiable subgroup (nine of ten environments), and under composition-OOD exchangeability breaks and overall coverage falls to 0.41–0.84 (Appendix Table 31).

6 Related work

Counterfactual world models. Zhou & Zhang (2026) share our setting and defer the supervision-matched control we run; PhyLatent (Zeng et al., 2026) and work on latent physical parameters (Zhang et al., 2026; Tan et al., 2026; Wang et al., 2026) label *parameters*, whereas we label *query answerability*; Counterfactual Quotient Models (Chen et al., 2026) quotient futures by what a decision needs, over actions rather than mechanisms. Causal benchmarks for video world models (Cai et al., 2026; Begiristain et al., 2026) score whether a generated future changes with an intervention, judged by a vision–language model or by consistency across renders; ours enumerates the answer.

Identifiability. The question descends from causal identification theory (Pearl, 2009; Shpitser & Pearl, 2008) and its partial-identification and bounding variants (Balke & Pearl, 1997; Bareinboim & Pearl, 2016); recent counterfactual identifiability theory (Chen & Du, 2025; Ribeiro et al., 2025; Rossetto & Antonucci, 2026; Raghavan & Bareinboim, 2026; Bynum et al., 2026) is likewise class-level and asymptotic, as is Bhadane et al. (2026)'s population-level test; ours is the per-episode, per-query verdict from finite evidence, computed exactly, which makes deciding whether to answer a selective-prediction problem (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) with known ground truth.

Decision-aware experimental design. Choosing evidence by expected information is classical (Lindley, 1956; Chaloner & Verdinelli, 1995) and now amortised into policies (Foster et al., 2019); GoBOED (Go et al., 2026), Causal-EPIG (Gao et al., 2026), RAED (Fotias, 2026) and others (Maringgele & Etesami, 2026; Hartmann et al., 2026) show that decision-targeted designs beat generic information gain. We supply the first setting in which the exact optimum is computable, so a learned policy is scored by regret rather than against another heuristic.

Baseline-sensitivity. Our loss finding joins results that dissolve under better-tuned baselines (Henderson et al., 2018; Lucic et al., 2018); a raw-unit loss handicapping the simpler model has not been reported for world models.

7 Discussion and limitations

On a benchmark with exact ground truth, the identifiability machinery of the design the field is converging on is not doing the work attributed to it. Under full supervision the learned posterior adds little (an exact posterior adds 0.015 AUROC on the ramp environments and less elsewhere), and the one apparent exception was an artifact of the baseline's training. The productive question is therefore not whether world models should carry causal structure but what supervision is available when they are trained: with abundant query-level supervision, explicit latent factorisation adds little; when it is scarce, a factorised posterior representation improves label efficiency, though it does not replace the label.

Two boundaries should be stated. The negative result concerns categorical posteriors over a finite bank; continuous latent-parameter designs are untested. The positive result lowers the label requirement, it does not remove it: with no labels both models sit at chance and the posterior's own dispersion reaches only 0.58–0.77 AUROC, so structure buys a hundredfold reduction, from thousands of labels to about forty, not label-free answerability. For baselines we recommend the tolerance-scaled loss, five seeds, a paired bootstrap with the ± 0.002 region, final-checkpoint selection when labels are scarce, and the supervision-matched control.

Limitations. Identifiability is bank-relative; two experiments bound the cost. With the bank a dense sample of one continuous four-term law (`spring_continuous`; an approximation, not exact inference) every conclusion repeats: equivalence at full supervision, $+0.004$–0.009 for structure at 0.1–5% of the labels, both agents recovering the oracle's 13% saving. With test episodes from a law outside the bank (Appendix Tables 24 and 15) the bank's verdict is wrong on 1–4% of queries, and on those both models assign $p(\text{identifiable})$ of 0.26–0.45 against 0.90–0.97 elsewhere (four of five systems): the head is not memorising bank-relative labels, though the off-bank laws are mild. The visual encoder is small. The sequential results use the benchmark's own probe menus, where the oracle's margin over generic gain is modest (7–31%).

REPRODUCIBILITY STATEMENT

All code, generation configuration and evaluation scripts are included in the supplementary material; the table aggregator refuses ensembles with fewer than five members and any run with missing seeds rather than averaging over them. The dataset is distributed as seeded, deterministic generation code rather than a frozen download. Every table and figure is regenerated from raw experiment artifacts by the included scripts. Appendix A gives the benchmark datasheet and hyper-parameters; the sequence of controls in Section 4, including the retracted claim, is preserved in the released audit log.

ETHICS STATEMENT

The work uses only synthetic physical and causal-simulation environments; no human subjects, personal data or sensitive content are involved. We see no direct path to harm. The benchmark's labels are exact relative to a fixed hypothesis bank, and we caution against reading identifiability verdicts obtained under one bank as statements about the physical world.

AI USE STATEMENT

AI was used in this work, under the authors' direction, for polish writing, improve scripts, find related work and generating synthetic dataset. The submission was also revised in response to the optional LLM feedback offered by the ICLR 2027 program chairs (Google's Paper Assistant Tool): the text-versus-table inconsistencies, missing definitions and additional controls it identified were checked by the authors and, where valid, corrected or run. All experimental claims were verified by the authors against the released artifacts. The authors take full responsibility for the content.

REFERENCES

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

León Begiristain, Olaf Dünkel, and Adam Kortylewski. CRONOS: Benchmarking counterfactual physical consistency in video models. *arXiv preprint arXiv:2605.23699*, 2026.

Sourbh Bhadane, Joris M. Mooij, Philip Boeken, and Onno Zoeter. Testing partially-identifiable causal queries using ternary tests. In *Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 549–566, 2026.

Lucius E. J. Bynum, Rajesh Ranganath, and Kyunghyun Cho. Computational identifiability. *arXiv preprint arXiv:2606.19361*, 2026.

Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. What-if world: A causal benchmark for general world models in embodied scenarios. *arXiv preprint arXiv:2605.27589*, 2026.

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

Junlin Chen, Ruijie Wang, and Jianxin Li. Counterfactual quotient models: Learning what actions change, not what the world does. *arXiv preprint arXiv:2608.22092*, 2026.

Yikang Chen and Dehui Du. Exogenous isomorphism for counterfactual identifiability. In *International Conference on Machine Learning*, 2025.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems*, 2019.

Sofianos Panagiotis Fotias. Resolution-aware experimental design under partial identifiability. *arXiv preprint arXiv:2609.03686*, 2026.

Erdun Gao, Jake Fawkes, and Dino Sejdinovic. Causal-EPIG: Causally aligned active CATE estimation. In *International Conference on Machine Learning*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Jinwoo Go, Xiaoning Qian, and Byung-Jun Yoon. Goal-driven Bayesian optimal experimental design for robust decision-making under model uncertainty. *arXiv preprint arXiv:2605.26093*, 2026.

Jakob Hartmann, James Harvey, Jhonathan Navott, Erik Y. Wang, Luckeciano C. Melo, Flaviu Cipcigan, Cheng Zhang, and Alessandro Abate. Amortising Bayesian experimental design for sequential information gathering in LLMs. In *ICML 2026 Workshop on Foundations of Deep Generative Models (FoGen)*, 2026. URL https://arxiv.org/abs/2607.03426.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

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

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems*, 2018.

Tobias Maringgele and Jalal Etesami. Optimal experiments for partial causal effect identification. *arXiv preprint arXiv:2605.06993*, 2026.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Arvind Raghavan and Elias Bareinboim. Causal identification from counterfactual data: Completeness and bounding results. In *International Conference on Machine Learning*, 2026.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. In *Advances in Neural Information Processing Systems*, 2025.

Eric Rossetto and Alessandro Antonucci. Partial identification under causal orders by linear programming. *arXiv preprint arXiv:2608.24427*, 2026.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Kaizhen Tan, Xin Xu, Siru Tao, Yixiao Li, Hanzhe Hong, Yang Feng, and Heqing Du. What can latent world models know? Physical parameter identifiability in multimodal predictive representations. *arXiv preprint arXiv:2607.27017*, 2026.

Rui Wang, Yeteng Wu, Xianlin Zhang, and Mengshi Qi. ExPhy: A benchmark for explicit physical property learning in multi-object trajectory forecasting. *arXiv preprint arXiv:2608.20009*, 2026.

Archana Warrier, Thanh Dat Nguyen, Michelangelo Naim, Moksh Jain, Yichao Liang, Karen Schroeder, Cambridge Yang, Joshua B. Tenenbaum, Sebastian Vollmer, Kevin Ellis, and Zenna Tavares. Benchmarking world-model learning with environment-level queries. In *International Conference on Machine Learning*, 2026.

Xi Zeng, Haojie Ren, and Ziying Song. PhyLatent: Learning dynamics-relevant representations for JEPA world models. *arXiv preprint arXiv:2608.05720*, 2026.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026.

Todd Y. Zhou and Daniel Zhang. Learning counterfactual world models for embodied reasoning under partial observability. *arXiv preprint arXiv:2609.05834*, 2026.

A BENCHMARK CHARACTERISATION AND DATASHEET

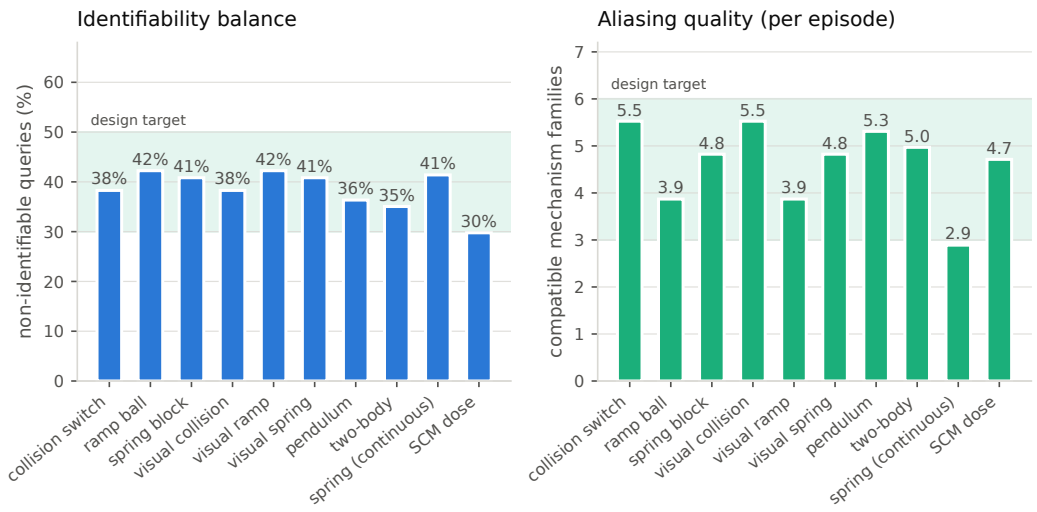

Figure 3: Identifiability balance and aliasing quality per environment, computed from the dataset with no model involved.

Per environment: 12,000 train / 1,500 validation / 2,000 test episodes per split, three queries per episode, evidence budget four, $\lambda_{\text{cost}} = 0.3$. Models: $d = 256$, 4 layers, 8 heads, AdamW (3×10^{-4}, weight decay 0.01, cosine schedule with 5% warm-up), 60 epochs, bf16, batch 512, model selection by validation identifiability AUROC at full supervision and the final checkpoint for every run below 100% labels (Section 3), temperature scaling on validation. Baselines share the encoder and budget. The full datasheet (Gebru et al. format) accompanies the code.

B FULL RESULT TABLES

Table 4: Sequential evidence acquisition on ambiguous IID queries (tolerance-scaled loss, 600 queries per split). "oracle"/"learned": probes to resolution with ground-truth stopping, so only action choices are scored; "vs oracle" is regret against the exact query-specific policy. The last three columns use the agent's own stopping rule. Full oracle-policy columns in Appendix Table 28.

env	model	seeds	oracle	learned	vs generic	vs oracle	self-stop probes	resolved	answer
collision	QID-AWM	5	1.76	1.78	-40.0%	+1.0%	3.14	0.73	0.92
collision	B1+ID	5	1.76	1.77	-40.3%	+0.4%	3.09	0.74	0.90
ramp	QID-AWM	5	1.03	1.05	-8.4%	+1.3%	3.64	0.93	0.51
ramp	B1+ID	5	1.03	1.04	-8.8%	+0.9%	3.36	0.89	0.50
spring	QID-AWM	5	1.06	1.10	-16.9%	+3.2%	3.54	0.93	0.78
spring	B1+ID	5	1.06	1.10	-16.8%	+3.3%	3.44	0.91	0.81
vis. collision	QID-AWM	3	1.74	1.76	-40.6%	+1.6%	3.25	0.75	0.91
vis. collision	B1+ID	3	1.74	1.76	-40.8%	+1.2%	3.17	0.74	0.91
vis. ramp	QID-AWM	3	1.03	1.06	-6.6%	+2.8%	3.63	0.92	0.47
vis. ramp	B1+ID	3	1.03	1.04	-7.7%	+1.6%	3.30	0.86	0.52
vis. spring	QID-AWM	3	1.06	1.11	-16.6%	+4.9%	3.58	0.91	0.69
vis. spring	B1+ID	3	1.06	1.10	-17.6%	+3.7%	3.28	0.87	0.68
pendulum	QID-AWM	5	1.31	1.40	+4.6%	+6.3%	3.71	0.94	0.64
pendulum	B1+ID	5	1.31	1.38	+3.3%	+5.0%	3.58	0.91	0.66
two-body	QID-AWM	5	1.26	1.38	-22.7%	+9.9%	3.79	0.89	0.72
two-body	B1+ID	5	1.26	1.37	-23.6%	+8.6%	3.73	0.88	0.71
spring (cont.)	QID-AWM	3	1.07	1.09	-13.0%	+2.2%	3.78	0.96	0.83
spring (cont.)	B1+ID	3	1.07	1.08	-13.8%	+1.3%	3.65	0.94	0.82
SCM dose	QID-AWM	3	1.11	1.19	-1.6%	+7.3%	3.75	0.93	0.46
SCM dose	B1+ID	3	1.11	1.18	-2.1%	+6.8%	3.62	0.91	0.46

Table 5: Learned mechanism posterior scored against the exact compatible set C (tolerance-scaled loss). Mass on compatible families (family-prior reference in brackets), top-1 compatible, recall at $|C|$, family-membership AUROC, Spearman correlation of predicted and exact dispersion on binary queries, and the AUROC of posterior entropy as an identifiability score.

env	seeds	mass on C (prior)	top-1 $\in C$	recall@$\|C\|$	membership AUROC	Spearman D_q	entropy AUROC (id.)
collision	5	0.991 (0.942)	0.997	0.919	0.809	0.83	0.331
ramp	5	0.749 (0.625)	0.825	0.696	0.667	0.43	0.478
spring	5	0.921 (0.793)	0.961	0.835	0.694	0.44	0.684
vis. collision	5	0.991 (0.943)	0.997	0.924	0.803	0.82	0.347
vis. ramp	5	0.746 (0.623)	0.824	0.708	0.667	0.42	0.502
vis. spring	5	0.909 (0.793)	0.952	0.831	0.676	0.44	0.689
pendulum	5	0.910 (0.860)	0.933	0.889	0.644	0.68	0.417
two-body	5	0.961 (0.868)	0.954	0.843	0.730	0.38	0.664
spring (cont.)	5	0.952 (0.742)	0.979	0.831	0.759	0.50	0.540
SCM dose	5	0.880 (0.812)	0.930	0.838	0.659	0.43	0.410

Table 6: B1+KHead at 1% of identifiability labels for $K \in \{2, 4, 6, 12, 24\}$, with the plain control and QID-AWM (IID AUROC, seeds as superscripts).

env	plain	K=2	K=4	K=6	K=12	K=24	QID-AWM
ramp	$0.946^{(10)}$	$0.949^{(5)}$	$0.950^{(5)}$	$0.950^{(8)}$	$0.950^{(5)}$	$0.949^{(5)}$	$0.961^{(5)}$
spring	$0.927^{(10)}$	$0.926^{(5)}$	$0.931^{(5)}$	$0.932^{(8)}$	$0.928^{(5)}$	$0.933^{(5)}$	$0.940^{(5)}$

Table 7: Checkpoint selection at scarce labels (IID AUROC, seeds as superscripts). The plain control has no selection. The structured model selected by validation identifiability AUROC on the fully labelled validation split, by validation counterfactual accuracy (no identifiability label), or taken at its final epoch — the protocol every scarce-label table in this paper reports.

env	labels	plain	val AUROC	val cf-acc	final epoch	Δ val-AUROC$-$plain	Δ final$-$plain
ramp	0.1%	$0.8861^{(10)}$	$0.9064^{(5)}$	$0.8967^{(5)}$	$0.8973^{(5)}$	+0.0190 $\neq$	+0.0099 $\neq$
ramp	1%	$0.9465^{(10)}$	$0.9606^{(5)}$	$0.9538^{(5)}$	$0.9549^{(5)}$	+0.0151 $\neq$	+0.0094 $\neq$
ramp	5%	$0.9605^{(5)}$	$0.9690^{(5)}$	–	$0.9640^{(5)}$	+0.0085 $\neq$	+0.0035 $\neq$
spring	0.1%	$0.8638^{(10)}$	$0.8858^{(5)}$	$0.8676^{(5)}$	$0.8609^{(5)}$	+0.0212 $\neq$	-0.0037 $\sim$
spring	1%	$0.9273^{(10)}$	$0.9402^{(5)}$	$0.9386^{(5)}$	$0.9398^{(5)}$	+0.0124 $\neq$	+0.0120 $\neq$
spring	5%	$0.9558^{(5)}$	$0.9636^{(5)}$	–	$0.9627^{(5)}$	+0.0078 $\neq$	+0.0069 $\neq$
vis. ramp	0.1%	$0.8939^{(5)}$	$0.9256^{(5)}$	–	$0.9120^{(5)}$	+0.0318 $\neq$	+0.0181 $\neq$
vis. ramp	1%	$0.9358^{(5)}$	$0.9569^{(5)}$	–	$0.9476^{(5)}$	+0.0211 $\neq$	+0.0119 $\neq$
vis. ramp	5%	$0.9545^{(5)}$	$0.9707^{(3)}$	–	$0.9597^{(5)}$	+0.0155 $\neq$	+0.0052 $\neq$
vis. spring	0.1%	$0.8550^{(5)}$	$0.8835^{(5)}$	–	$0.8499^{(5)}$	+0.0285 $\neq$	-0.0051 $\sim$
vis. spring	1%	$0.9258^{(5)}$	$0.9355^{(5)}$	–	$0.9384^{(5)}$	+0.0096 $\neq$	+0.0125 $\neq$
vis. spring	5%	$0.9452^{(5)}$	$0.9556^{(3)}$	–	$0.9555^{(5)}$	+0.0097 $\neq$	+0.0103 $\neq$
pendulum	0.1%	$0.8180^{(10)}$	$0.8843^{(5)}$	$0.8660^{(5)}$	$0.8563^{(5)}$	+0.0636 $\neq$	+0.0356 $\neq$
pendulum	1%	$0.9607^{(10)}$	$0.9669^{(5)}$	$0.9659^{(5)}$	$0.9648^{(5)}$	+0.0052 $\neq$	+0.0032 $\neq$
pendulum	5%	$0.9765^{(5)}$	$0.9801^{(5)}$	–	$0.9807^{(5)}$	+0.0036 $\neq$	+0.0042 $\neq$
two-body	0.1%	$0.9310^{(10)}$	$0.9451^{(5)}$	$0.9416^{(5)}$	$0.9437^{(5)}$	+0.0148 $\neq$	+0.0134 $\neq$
two-body	1%	$0.9632^{(10)}$	$0.9734^{(5)}$	$0.9748^{(5)}$	$0.9743^{(5)}$	+0.0105 $\neq$	+0.0114 $\neq$
two-body	5%	$0.9873^{(5)}$	$0.9906^{(5)}$	–	$0.9903^{(5)}$	+0.0033 $\neq$	+0.0031 $\neq$
spring (cont.)	0.1%	$0.9025^{(5)}$	$0.9080^{(5)}$	–	$0.9115^{(5)}$	+0.0056 $\sim$	+0.0090 $\neq$
spring (cont.)	1%	$0.9556^{(5)}$	$0.9629^{(5)}$	–	$0.9628^{(5)}$	+0.0074 $\neq$	+0.0072 $\neq$
spring (cont.)	5%	$0.9709^{(5)}$	$0.9787^{(5)}$	–	$0.9745^{(5)}$	+0.0079 $\neq$	+0.0037 $\neq$
SCM dose	0.1%	$0.9120^{(5)}$	$0.9482^{(5)}$	–	$0.9362^{(5)}$	+0.0362 $\neq$	+0.0243 $\neq$
SCM dose	1%	$0.9515^{(5)}$	$0.9695^{(5)}$	–	$0.9729^{(5)}$	+0.0180 $\neq$	+0.0214 $\neq$
SCM dose	5%	$0.9810^{(5)}$	$0.9886^{(5)}$	–	$0.9884^{(5)}$	+0.0076 $\neq$	+0.0073 $\neq$

Table 8: The two mechanism controls (tolerance-scaled loss, IID AUROC, seeds as superscripts, final checkpoints). *KHead+dc*: B1+KHead trained with QID-AWM's dispersion and calibration losses on its gate mixture; *QID-AWM−dc*: the structured model without those two losses. Verdicts are paired-bootstrap.

env	labels	plain	KHead	KHead+dc	QID-AWM	QID-AWM−dc	Δ KHead+dc−plain	Δ (−dc)−QID-AWM
ramp	0.1%	$0.8861^{(10)}$	$0.8895^{(8)}$	$0.8994^{(5)}$	$0.8973^{(5)}$	$0.8934^{(5)}$	+0.0120 ≠	-0.0039 ∼
ramp	1%	$0.9465^{(10)}$	$0.9499^{(8)}$	$0.9491^{(5)}$	$0.9549^{(5)}$	$0.9533^{(5)}$	+0.0036 ∼	-0.0016 ∼
spring	0.1%	$0.8638^{(10)}$	$0.8568^{(8)}$	$0.8598^{(5)}$	$0.8609^{(5)}$	$0.8690^{(5)}$	-0.0048 ∼	+0.0080 ∼
spring	1%	$0.9273^{(10)}$	$0.9322^{(8)}$	$0.9282^{(5)}$	$0.9398^{(5)}$	$0.9430^{(5)}$	+0.0004 ∼	+0.0032 ∼
pendulum	0.1%	$0.8180^{(10)}$	$0.8232^{(8)}$	$0.8232^{(5)}$	$0.8563^{(5)}$	$0.8675^{(5)}$	+0.0026 ∼	+0.0112 ≠
pendulum	1%	$0.9607^{(10)}$	$0.9613^{(8)}$	$0.9598^{(5)}$	$0.9648^{(5)}$	$0.9650^{(5)}$	-0.0019 ≠	+0.0001 ∼
two-body	0.1%	$0.9310^{(10)}$	$0.9363^{(8)}$	$0.9341^{(5)}$	$0.9437^{(5)}$	$0.9441^{(5)}$	+0.0038 ∼	+0.0004 ∼
two-body	1%	$0.9632^{(10)}$	$0.9643^{(8)}$	$0.9631^{(5)}$	$0.9743^{(5)}$	$0.9742^{(5)}$	+0.0002 ∼	-0.0001 ∼
spring (cont.)	0.1%	$0.9025^{(5)}$	$0.9031^{(5)}$	$0.9053^{(5)}$	$0.9115^{(5)}$	$0.9058^{(5)}$	+0.0028 ∼	-0.0057 ∼
spring (cont.)	1%	$0.9556^{(5)}$	$0.9563^{(5)}$	$0.9571^{(5)}$	$0.9628^{(5)}$	$0.9620^{(5)}$	+0.0015 ∼	-0.0008 ∼
SCM dose	0.1%	$0.9120^{(5)}$	$0.9101^{(5)}$	$0.9219^{(5)}$	$0.9362^{(5)}$	$0.9301^{(5)}$	+0.0099 ≠	-0.0061 ∼
SCM dose	1%	$0.9515^{(5)}$	$0.9517^{(5)}$	$0.9523^{(5)}$	$0.9729^{(5)}$	$0.9662^{(5)}$	+0.0008 ∼	-0.0067 ≠

Table 9: Query-type shortcut check. Share of IID test queries whose query type has a training base rate of identifiability in $[0.3, 0.7]$ ("balanced"), and identifiability AUROC of the query-type prior B0 and of both models on all queries and on the balanced subset (tolerance-scaled loss, five seeds).

env	balanced share	B0 (all)	QID-AWM (all)	B1+ID (all)	B0 (bal.)	QID-AWM (bal.)	B1+ID (bal.)
collision	68%	0.741	1.000	1.000	0.589	1.000	1.000
ramp	21%	0.883	0.981	0.981	0.627	0.951	0.951
spring	16%	0.868	0.984	0.984	0.526	0.951	0.951
vis. collision	68%	0.741	1.000	1.000	0.589	1.000	1.000
vis. ramp	21%	0.883	0.977	0.977	0.627	0.947	0.945
vis. spring	16%	0.868	0.977	0.972	0.526	0.932	0.915
pendulum	50%	0.810	0.992	0.992	0.596	0.989	0.990
two-body	30%	0.885	0.997	0.997	0.559	0.996	0.997
spring (cont.)	17%	0.895	0.988	0.988	0.678	0.982	0.983
SCM dose	0%	0.949	0.995	0.994	–	–	–

Table 10: Oracle-posterior upper bound (tolerance-scaled loss, IID AUROC, seeds as superscripts). Scarce-label rows: both models at their final checkpoint. +*oracle post.*: QID-AWM with the exact compatible family set from the bank in place of its learned posterior (uniform over compatible families, not over compatible grid points; family loss off). Δ is oracle minus QID-AWM with its paired-bootstrap verdict.

env	labels	plain	QID-AWM	+oracle post.	Δ oracle−QID-AWM
ramp	100%	$0.9807^{(5)}$	$0.9806^{(5)}$	$0.9947^{(5)}$	+0.0141 ≠
ramp	1%	$0.9465^{(10)}$	$0.9549^{(5)}$	$0.9571^{(5)}$	+0.0022 ∼
ramp	0.1%	$0.8861^{(10)}$	$0.8973^{(5)}$	$0.9085^{(5)}$	+0.0113 ≠
spring	100%	$0.9840^{(5)}$	$0.9836^{(5)}$	$0.9867^{(5)}$	+0.0031 ≠
spring	1%	$0.9273^{(10)}$	$0.9398^{(5)}$	$0.9353^{(5)}$	-0.0045 ≠
spring	0.1%	$0.8638^{(10)}$	$0.8609^{(5)}$	$0.8531^{(5)}$	-0.0078 ∼
vis. ramp	100%	$0.9768^{(10)}$	$0.9776^{(10)}$	$0.9921^{(3)}$	+0.0147 ≠
vis. spring	100%	$0.9721^{(5)}$	$0.9772^{(5)}$	$0.9805^{(3)}$	+0.0038 ≠
pendulum	100%	$0.9924^{(5)}$	$0.9922^{(5)}$	$0.9931^{(5)}$	+0.0009 ≡
pendulum	1%	$0.9607^{(10)}$	$0.9648^{(5)}$	$0.9641^{(5)}$	-0.0007 ∼
pendulum	0.1%	$0.8180^{(10)}$	$0.8563^{(5)}$	$0.8817^{(5)}$	+0.0255 ≠
two-body	100%	$0.9970^{(5)}$	$0.9970^{(5)}$	$0.9967^{(5)}$	-0.0003 ≡
two-body	1%	$0.9632^{(10)}$	$0.9743^{(5)}$	$0.9713^{(5)}$	-0.0030 ≠
two-body	0.1%	$0.9310^{(10)}$	$0.9437^{(5)}$	$0.9483^{(5)}$	+0.0046 ≠
spring (cont.)	100%	$0.9876^{(5)}$	$0.9878^{(5)}$	$0.9913^{(5)}$	+0.0036 ≠
spring (cont.)	1%	$0.9556^{(5)}$	$0.9628^{(5)}$	$0.9650^{(5)}$	+0.0022 ∼
spring (cont.)	0.1%	$0.9025^{(5)}$	$0.9115^{(5)}$	$0.9301^{(5)}$	+0.0186 ≠
SCM dose	100%	$0.9943^{(5)}$	$0.9949^{(5)}$	$0.9942^{(5)}$	-0.0007 ≡
SCM dose	1%	$0.9515^{(5)}$	$0.9729^{(5)}$	$0.9664^{(5)}$	-0.0065 ≠
SCM dose	0.1%	$0.9120^{(5)}$	$0.9362^{(5)}$	$0.9401^{(5)}$	+0.0038 ∼

Table 11: B1+KHead with the gate's disagreement statistics routed to the identifiability head (*KHead+gate*) and with one independent counterfactual head per expert (*KHead-ind*) against KHead, the plain control and QID-AWM at 0.1% and 1% of identifiability labels (tolerance-scaled loss, IID AUROC, seeds as superscripts, final checkpoints). The two Δ columns are KHead+gate and KHead-ind minus QID-AWM, with verdicts.

env	labels	plain	KHead	KHead+gate	KHead-ind	QID-AWM	Δ gate$-$QID-AWM	Δ ind$-$QID-AWM
ramp	0.1%	$0.8861^{(10)}$	$0.8895^{(8)}$	$0.8945^{(5)}$	$0.8855^{(5)}$	$0.8973^{(5)}$	-0.0028 $\sim$	-0.0118 $\neq$
ramp	1%	$0.9465^{(10)}$	$0.9499^{(8)}$	$0.9494^{(5)}$	$0.9502^{(5)}$	$0.9549^{(5)}$	-0.0055 $\neq$	-0.0047 $\neq$
spring	0.1%	$0.8638^{(10)}$	$0.8568^{(8)}$	$0.8568^{(5)}$	$0.8578^{(5)}$	$0.8609^{(5)}$	-0.0042 $\sim$	-0.0032 $\sim$
spring	1%	$0.9273^{(10)}$	$0.9322^{(8)}$	$0.9265^{(5)}$	$0.9305^{(5)}$	$0.9398^{(5)}$	-0.0133 $\neq$	-0.0093 $\neq$
vis. ramp	0.1%	$0.8939^{(5)}$	$0.8979^{(5)}$	$0.8982^{(5)}$	$0.8940^{(5)}$	$0.9120^{(5)}$	-0.0138 $\neq$	-0.0180 $\neq$
vis. ramp	1%	$0.9358^{(5)}$	$0.9381^{(5)}$	$0.9379^{(5)}$	$0.9372^{(5)}$	$0.9476^{(5)}$	-0.0098 $\neq$	-0.0105 $\neq$
vis. spring	0.1%	$0.8550^{(5)}$	$0.8482^{(5)}$	$0.8508^{(5)}$	$0.8575^{(5)}$	$0.8499^{(5)}$	+0.0010 $\sim$	+0.0076 $\neq$
vis. spring	1%	$0.9258^{(5)}$	$0.9263^{(5)}$	$0.9311^{(5)}$	$0.9335^{(5)}$	$0.9384^{(5)}$	-0.0073 $\neq$	-0.0048 $\neq$
pendulum	0.1%	$0.8180^{(10)}$	$0.8232^{(8)}$	$0.8330^{(5)}$	$0.8169^{(5)}$	$0.8563^{(5)}$	-0.0232 $\neq$	-0.0393 $\neq$
pendulum	1%	$0.9607^{(10)}$	$0.9613^{(8)}$	$0.9622^{(5)}$	$0.9587^{(5)}$	$0.9648^{(5)}$	-0.0027 $\neq$	-0.0061 $\neq$
two-body	0.1%	$0.9310^{(10)}$	$0.9363^{(8)}$	$0.9348^{(5)}$	$0.9407^{(5)}$	$0.9437^{(5)}$	-0.0089 $\neq$	-0.0030 $\neq$
two-body	1%	$0.9632^{(10)}$	$0.9643^{(8)}$	$0.9652^{(5)}$	$0.9672^{(5)}$	$0.9743^{(5)}$	-0.0091 $\neq$	-0.0071 $\neq$
spring (cont.)	0.1%	$0.9025^{(5)}$	$0.9031^{(5)}$	$0.9094^{(5)}$	$0.9107^{(5)}$	$0.9115^{(5)}$	-0.0021 $\sim$	-0.0008 $\sim$
spring (cont.)	1%	$0.9556^{(5)}$	$0.9563^{(5)}$	$0.9508^{(5)}$	$0.9578^{(5)}$	$0.9628^{(5)}$	-0.0120 $\neq$	-0.0050 $\neq$
SCM dose	0.1%	$0.9120^{(5)}$	$0.9101^{(5)}$	$0.9176^{(5)}$	$0.9146^{(5)}$	$0.9362^{(5)}$	-0.0187 $\neq$	-0.0216 $\neq$
SCM dose	1%	$0.9515^{(5)}$	$0.9517^{(5)}$	$0.9521^{(5)}$	$0.9501^{(5)}$	$0.9729^{(5)}$	-0.0207 $\neq$	-0.0228 $\neq$

Table 12: Plain control at twice the width (d=512, four layers) in the scarce-label regime, against the standard control (d=256) and QID-AWM (tolerance-scaled loss, IID AUROC, seeds as superscripts). Structured runs at their final checkpoint. Δ is QID-AWM minus the wide control with its verdict.

env	labels	plain d=256	plain d=512	QID-AWM	Δ QID-AWM$-$wide
ramp	0.1%	$0.8861^{(10)}$	$0.8919^{(5)}$	$0.8973^{(5)}$	+0.0054 $\sim$
ramp	1%	$0.9465^{(10)}$	$0.9513^{(5)}$	$0.9549^{(5)}$	+0.0036 $\neq$
spring	0.1%	$0.8638^{(10)}$	$0.8537^{(5)}$	$0.8609^{(5)}$	+0.0073 $\sim$
spring	1%	$0.9273^{(10)}$	$0.9319^{(5)}$	$0.9398^{(5)}$	+0.0079 $\neq$
pendulum	0.1%	$0.8180^{(10)}$	$0.8330^{(5)}$	$0.8563^{(5)}$	+0.0233 $\sim$
pendulum	1%	$0.9607^{(10)}$	$0.9613^{(5)}$	$0.9648^{(5)}$	+0.0035 $\neq$
two-body	0.1%	$0.9310^{(10)}$	$0.9383^{(5)}$	$0.9437^{(5)}$	+0.0055 $\neq$
two-body	1%	$0.9632^{(10)}$	$0.9676^{(5)}$	$0.9743^{(5)}$	+0.0067 $\neq$
spring (cont.)	0.1%	$0.9025^{(5)}$	$0.8991^{(5)}$	$0.9115^{(5)}$	+0.0124 $\neq$
spring (cont.)	1%	$0.9556^{(5)}$	$0.9529^{(5)}$	$0.9628^{(5)}$	+0.0099 $\neq$
SCM dose	0.1%	$0.9120^{(5)}$	$0.9202^{(5)}$	$0.9362^{(5)}$	+0.0160 $\neq$
SCM dose	1%	$0.9515^{(5)}$	$0.9506^{(5)}$	$0.9729^{(5)}$	+0.0222 $\neq$

Table 13: Scarce-label margin over the plain control for the full structured model and for the variant without the mechanism-family loss (*no-mech*), IID AUROC, seeds as superscripts, with paired-bootstrap verdicts. Structured runs at their final checkpoint.

	0.1% labels		1% labels		5% labels	
env	QID-AWM	no-mech	QID-AWM	no-mech	QID-AWM	no-mech
ramp	+0.0099$^{(5)}$ $\neq$	+0.0214$^{(5)}$ $\neq$	+0.0094$^{(5)}$ $\neq$	+0.0084$^{(5)}$ $\neq$	+0.0035$^{(5)}$ $\neq$	+0.0040$^{(5)}$ $\neq$
spring	-0.0037$^{(5)}$ $\sim$	-0.0057$^{(5)}$ $\sim$	+0.0120$^{(5)}$ $\neq$	+0.0091$^{(5)}$ $\neq$	+0.0069$^{(5)}$ $\neq$	+0.0051$^{(5)}$ $\neq$
vis. ramp	+0.0181$^{(5)}$ $\neq$	+0.0247$^{(5)}$ $\neq$	+0.0119$^{(5)}$ $\neq$	+0.0107$^{(5)}$ $\neq$	+0.0052$^{(5)}$ $\neq$	+0.0043$^{(5)}$ $\neq$
vis. spring	-0.0051$^{(5)}$ $\sim$	+0.0019$^{(5)}$ $\sim$	+0.0125$^{(5)}$ $\neq$	+0.0116$^{(5)}$ $\neq$	+0.0103$^{(5)}$ $\neq$	+0.0081$^{(5)}$ $\neq$
pendulum	+0.0356$^{(5)}$ $\neq$	+0.0650$^{(5)}$ $\neq$	+0.0032$^{(5)}$ $\neq$	+0.0039$^{(5)}$ $\neq$	+0.0042$^{(5)}$ $\neq$	+0.0039$^{(5)}$ $\neq$
two-body	+0.0134$^{(5)}$ $\neq$	+0.0172$^{(5)}$ $\neq$	+0.0114$^{(5)}$ $\neq$	+0.0084$^{(5)}$ $\neq$	+0.0031$^{(5)}$ $\neq$	+0.0007$^{(5)}$ $\equiv$
spring (cont.)	+0.0090$^{(5)}$ $\neq$	+0.0238$^{(5)}$ $\neq$	+0.0072$^{(5)}$ $\neq$	+0.0087$^{(5)}$ $\neq$	+0.0037$^{(5)}$ $\neq$	+0.0031$^{(5)}$ $\neq$
SCM dose	+0.0243$^{(5)}$ $\neq$	+0.0278$^{(5)}$ $\neq$	+0.0214$^{(5)}$ $\neq$	+0.0120$^{(5)}$ $\neq$	+0.0073$^{(5)}$ $\neq$	+0.0030$^{(5)}$ $\neq$

Table 14: Absolute identifiability AUROC across observation noise, QID-AWM / B1+ID (tolerance-scaled loss, five seeds), for the two environments whose margin depends on noise. On the state system both models degrade and the control degrades faster; from pixels both plateau and the margin closes.

env	noise 0	noise 0.15	noise 0.3	noise 0.5	noise 1
spring	0.984 / 0.984	0.979 / 0.976	0.973 / 0.968	0.965 / 0.958	0.950 / 0.941
vis. spring	0.977 / 0.972	0.970 / 0.965	0.969 / 0.965	0.968 / 0.967	0.966 / 0.966

Table 15: When the bank is wrong. Mean $p(\text{identifiable})$ of each model on out-of-bank test queries the training bank labels correctly ("bank right") and on the silent ones it labels identifiable although reality disagrees ("bank wrong"), with the silent fraction; tolerance-scaled loss, five seeds. Both models are markedly less confident exactly where the bank is wrong, on four of five systems.

env	QID-AWM: bank right	bank wrong	B1+ID: bank right	bank wrong	silent
ramp	0.90	0.35	0.92	0.40	2%
spring	0.93	0.34	0.94	0.39	4%
pendulum	0.94	0.40	0.95	0.45	1%
two-body	0.96	0.92	0.97	0.94	1%
SCM dose	0.96	0.26	0.97	0.30	2%

Table 16: Joint supervision reduction (tolerance-scaled loss, IID AUROC, five seeds). The same subset of training queries carries the identifiability label, the counterfactual-answer target and the evidence-gain target; every other query carries none of them, for both models. Cells: QID-AWM minus B1+ID with the three-way verdict.

env	1%	5%	25%
ramp	$+0.0051^{(5)}$ $\sim$	$+0.0010^{(5)}$ $\sim$	$+0.0003^{(5)}$ $\equiv$
spring	$+0.0021^{(5)}$ $\sim$	$+0.0102^{(5)}$ $\neq$	$+0.0024^{(5)}$ $\neq$
pendulum	$+0.0006^{(5)}$ $\sim$	$+0.0015^{(5)}$ $\sim$	$+0.0004^{(5)}$ $\equiv$

Table 17: Supervision-matched comparison, IID split, both training losses: identifiability AUROC, mean±std over seeds; Δ is QID-AWM minus B1+ID with its 95% paired-bootstrap interval, seeds as superscript, and the verdict ($\equiv$ within ± 0.002, $\neq$ excludes zero, $\sim$ otherwise).

	raw-unit scalar loss			tolerance-scaled scalar loss		
env	QID-AWM	B1+ID	Δ [95% CI]	QID-AWM	B1+ID	Δ [95% CI]
collision	1.000±0.000	1.000±0.000	+0.0001 [+0.0000, +0.0001]$^{(5)}$ $\equiv$	1.000±0.000	1.000±0.000	+0.0001 [+0.0000, +0.0001]$^{(5)}$ $\equiv$
ramp	0.993±0.003	0.984±0.001	+0.0094 [+0.0080, +0.0108]$^{(5)}$ $\neq$	0.981±0.000	0.981±0.000	-0.0001 [-0.0003, +0.0002]$^{(5)}$ $\equiv$
spring	0.983±0.000	0.983±0.000	-0.0001 [-0.0006, +0.0005]$^{(5)}$ $\equiv$	0.984±0.000	0.984±0.000	-0.0004 [-0.0006, -0.0001]$^{(5)}$ $\equiv$
vis. collision	1.000±0.000	1.000±0.000	+0.0001 [+0.0000, +0.0001]$^{(5)}$ $\equiv$	1.000±0.000	1.000±0.000	+0.0000 [-0.0000, +0.0001]$^{(5)}$ $\equiv$
vis. ramp	0.977±0.001	0.968±0.001	+0.0088 [+0.0079, +0.0097]$^{(5)}$ $\neq$	0.978±0.000	0.977±0.000	+0.0008 [+0.0004, +0.0011]$^{(10)}$ $\equiv$
vis. spring	0.976±0.000	0.967±0.001	+0.0097 [+0.0087, +0.0108]$^{(5)}$ $\neq$	0.977±0.001	0.972±0.001	+0.0051 [+0.0035, +0.0065]$^{(5)}$ $\neq$
pendulum	0.993±0.000	0.993±0.000	-0.0002 [-0.0005, +0.0001]$^{(5)}$ $\equiv$	0.992±0.000	0.992±0.000	-0.0003 [-0.0005, -0.0001]$^{(5)}$ $\equiv$
two-body	0.997±0.000	0.997±0.000	-0.0001 [-0.0004, +0.0001]$^{(5)}$ $\equiv$	0.997±0.000	0.997±0.000	+0.0000 [-0.0003, +0.0003]$^{(5)}$ $\equiv$
spring (cont.)	0.988±0.000	0.987±0.000	+0.0010 [+0.0003, +0.0017]$^{(5)}$ $\equiv$	0.988±0.000	0.988±0.000	+0.0002 [-0.0002, +0.0006]$^{(5)}$ $\equiv$
SCM dose	0.996±0.000	0.996±0.000	+0.0000 [-0.0003, +0.0004]$^{(5)}$ $\equiv$	0.995±0.000	0.994±0.000	+0.0005 [+0.0001, +0.0011]$^{(5)}$ $\equiv$

Table 18: Identifiability AUROC of QID-AWM / B1+ID on every split (tolerance-scaled loss). Parameter-OOD holds out a range of one control variable, mechanism-OOD one family, composition-OOD (family, intervention) pairs seen only separately in training.

env	seeds	IID	param.	mech.	comp.
collision	5	1.000 / 1.000	0.999 / 0.999	1.000 / 1.000	1.000 / 1.000
ramp	5	0.981 / 0.981	0.975 / 0.975	0.981 / 0.981	0.958 / 0.958
spring	5	0.984 / 0.984	0.982 / 0.981	0.987 / 0.987	0.979 / 0.979
vis. collision	5	1.000 / 1.000	1.000 / 1.000	1.000 / 1.000	1.000 / 1.000
vis. ramp	10	0.978 / 0.977	0.923 / 0.898	0.978 / 0.975	0.955 / 0.947
vis. spring	5	0.977 / 0.972	0.973 / 0.966	0.981 / 0.979	0.973 / 0.969
pendulum	5	0.992 / 0.992	0.993 / 0.994	0.991 / 0.991	0.988 / 0.988
two-body	5	0.997 / 0.997	0.994 / 0.994	0.996 / 0.996	0.997 / 0.997
spring (cont.)	5	0.988 / 0.988	0.980 / 0.980	0.967 / 0.968	0.988 / 0.988
SCM dose	5	0.995 / 0.994	0.993 / 0.992	0.992 / 0.992	0.982 / 0.979

Table 19: Appearance shift (identical physics, seeds and labels; re-rendered pixels) under the tolerance-scaled loss, five seeds. Both models keep their identifiability AUROC within 0.07 and lose a third of their counterfactual accuracy on the visual ramp: the answer head, not the identifiability head, is appearance-sensitive.

env	model	seeds	AUROC IID / shifted	Δ AUROC	ECE IID / shifted	cf-acc IID / shifted	Δ cf-acc
vis. collision	QID-AWM	5	0.9996 / 0.9943	-0.0053	0.005 / 0.054	0.939 / 0.914	-0.025
vis. collision	B1+ID	5	0.9996 / 0.9890	-0.0106	0.006 / 0.064	0.938 / 0.920	-0.018
vis. ramp	QID-AWM	5	0.9774 / 0.9344	-0.0430	0.030 / 0.086	0.809 / 0.471	-0.338
vis. ramp	B1+ID	5	0.9768 / 0.9092	-0.0676	0.026 / 0.128	0.823 / 0.442	-0.380
vis. spring	QID-AWM	5	0.9772 / 0.9534	-0.0239	0.023 / 0.037	0.913 / 0.806	-0.106
vis. spring	B1+ID	5	0.9721 / 0.9508	-0.0213	0.059 / 0.092	0.913 / 0.834	-0.079

Table 20: Label efficiency, full sweep (tolerance-scaled loss, IID AUROC). Cells: QID-AWM minus B1+ID, seeds as superscript; $\neq$ a paired-bootstrap CI excluding zero, $\equiv$ a CI inside ± 0.002. 100% is the main run.

env	0.1%	0.5%	1%	5%	10%	25%	100%
ramp	$+0.0099^{(5)} \neq$	$+0.0230^{(5)} \neq$	$+0.0094^{(5)} \neq$	$+0.0035^{(5)} \neq$	$+0.0023^{(5)} \neq$	$+0.0007^{(5)} \equiv$	$-0.0001^{(5)} \equiv$
spring	$-0.0037^{(5)} \sim$	$+0.0165^{(5)} \neq$	$+0.0120^{(5)} \neq$	$+0.0069^{(5)} \neq$	$+0.0041^{(5)} \neq$	$+0.0018^{(5)} \neq$	$-0.0004^{(5)} \equiv$
vis. ramp	$+0.0181^{(5)} \neq$	$+0.0248^{(5)} \neq$	$+0.0119^{(5)} \neq$	$+0.0052^{(5)} \neq$	$+0.0046^{(5)} \neq$	$+0.0014^{(5)} \neq$	$+0.0008^{(10)} \equiv$
vis. spring	$-0.0051^{(5)} \sim$	$+0.0168^{(5)} \neq$	$+0.0125^{(5)} \neq$	$+0.0103^{(5)} \neq$	$+0.0033^{(5)} \neq$	$+0.0000^{(5)} \equiv$	$+0.0051^{(5)} \neq$
pendulum	$+0.0356^{(5)} \neq$	$+0.0040^{(5)} \neq$	$+0.0032^{(5)} \neq$	$+0.0042^{(5)} \neq$	$+0.0015^{(5)} \neq$	$+0.0008^{(5)} \equiv$	$-0.0003^{(5)} \equiv$
two-body	$+0.0134^{(5)} \neq$	$+0.0135^{(5)} \neq$	$+0.0114^{(5)} \neq$	$+0.0031^{(5)} \neq$	$+0.0029^{(5)} \neq$	$+0.0010^{(5)} \equiv$	$+0.0000^{(5)} \equiv$
spring (cont.)	$+0.0090^{(5)} \neq$	$+0.0104^{(5)} \neq$	$+0.0072^{(5)} \neq$	$+0.0037^{(5)} \neq$	$+0.0019^{(5)} \neq$	$+0.0016^{(5)} \equiv$	$+0.0002^{(5)} \equiv$
SCM dose	$+0.0243^{(5)} \neq$	$+0.0160^{(5)} \neq$	$+0.0214^{(5)} \neq$	$+0.0073^{(5)} \neq$	$+0.0040^{(5)} \neq$	$+0.0017^{(5)} \neq$	$+0.0005^{(5)} \equiv$

Table 21: The six-model matrix at 0.1, 1, 5 and 100% of identifiability labels (absolute IID AUROC, seeds as superscripts): the plain control, the control with the auxiliary mechanism-family head (+MechAux), the control with a generic K-expert head (B1+KHead), the structured model without mechanism supervision (no mech.), with scrambled mechanism categories (scrambled mech.), and the full structured model.

env	model	0.1%	1%	5%	100%
ramp	plain	$0.8861^{(10)}$	$0.9465^{(10)}$	$0.9605^{(5)}$	$0.9807^{(5)}$
ramp	+MechAux	$0.8724^{(5)}$	$0.9456^{(5)}$	$0.9600^{(3)}$	$0.9806^{(3)}$
ramp	QID-AWM (no mech.)	$0.9088^{(5)}$	$0.9539^{(5)}$	$0.9645^{(5)}$	$0.9810^{(3)}$
ramp	QID-AWM (scrambled mech.)	$0.9066^{(5)}$	$0.9520^{(5)}$	$0.9650^{(5)}$	–
ramp	B1+KHead	$0.8895^{(8)}$	$0.9499^{(8)}$	$0.9597^{(3)}$	–
ramp	QID-AWM	$0.8973^{(5)}$	$0.9549^{(5)}$	$0.9640^{(5)}$	$0.9806^{(5)}$
spring	plain	$0.8638^{(10)}$	$0.9273^{(10)}$	$0.9558^{(5)}$	$0.9840^{(5)}$
spring	+MechAux	$0.8702^{(5)}$	$0.9298^{(5)}$	$0.9575^{(3)}$	$0.9837^{(3)}$
spring	QID-AWM (no mech.)	$0.8589^{(5)}$	$0.9369^{(5)}$	$0.9609^{(5)}$	$0.9840^{(3)}$
spring	QID-AWM (scrambled mech.)	$0.8576^{(5)}$	$0.9367^{(5)}$	$0.9605^{(5)}$	–
spring	B1+KHead	$0.8568^{(8)}$	$0.9322^{(8)}$	$0.9553^{(3)}$	–
spring	QID-AWM	$0.8609^{(5)}$	$0.9398^{(5)}$	$0.9627^{(5)}$	$0.9836^{(5)}$
vis. ramp	plain	$0.8939^{(5)}$	$0.9358^{(5)}$	$0.9545^{(5)}$	$0.9768^{(10)}$
vis. ramp	+MechAux	$0.8932^{(5)}$	$0.9357^{(5)}$	$0.9541^{(5)}$	$0.9767^{(3)}$
vis. ramp	QID-AWM (no mech.)	$0.9186^{(5)}$	$0.9465^{(5)}$	$0.9588^{(5)}$	$0.9783^{(3)}$
vis. ramp	QID-AWM (scrambled mech.)	$0.9213^{(5)}$	$0.9456^{(5)}$	$0.9590^{(5)}$	–
vis. ramp	B1+KHead	$0.8979^{(5)}$	$0.9381^{(5)}$	–	–
vis. ramp	QID-AWM	$0.9120^{(5)}$	$0.9476^{(5)}$	$0.9597^{(5)}$	$0.9776^{(10)}$
vis. spring	plain	$0.8550^{(5)}$	$0.9258^{(5)}$	$0.9452^{(5)}$	$0.9721^{(5)}$
vis. spring	+MechAux	$0.8545^{(5)}$	$0.9286^{(5)}$	$0.9503^{(5)}$	$0.9695^{(3)}$
vis. spring	QID-AWM (no mech.)	$0.8568^{(5)}$	$0.9374^{(5)}$	$0.9533^{(5)}$	$0.9776^{(3)}$
vis. spring	QID-AWM (scrambled mech.)	$0.8427^{(5)}$	$0.9368^{(5)}$	$0.9533^{(5)}$	–
vis. spring	B1+KHead	$0.8482^{(5)}$	$0.9263^{(5)}$	–	–
vis. spring	QID-AWM	$0.8499^{(5)}$	$0.9384^{(5)}$	$0.9555^{(5)}$	$0.9772^{(5)}$
pendulum	plain	$0.8180^{(10)}$	$0.9607^{(10)}$	$0.9765^{(5)}$	$0.9924^{(5)}$
pendulum	+MechAux	$0.8298^{(5)}$	$0.9609^{(5)}$	$0.9763^{(3)}$	$0.9927^{(3)}$
pendulum	QID-AWM (no mech.)	$0.8856^{(5)}$	$0.9655^{(5)}$	$0.9804^{(5)}$	$0.9918^{(3)}$
pendulum	QID-AWM (scrambled mech.)	$0.8775^{(5)}$	$0.9655^{(5)}$	$0.9794^{(5)}$	–
pendulum	B1+KHead	$0.8232^{(8)}$	$0.9613^{(8)}$	$0.9767^{(3)}$	–
pendulum	QID-AWM	$0.8563^{(5)}$	$0.9648^{(5)}$	$0.9807^{(5)}$	$0.9922^{(5)}$
two-body	plain	$0.9310^{(10)}$	$0.9632^{(10)}$	$0.9873^{(5)}$	$0.9970^{(5)}$
two-body	+MechAux	$0.9333^{(5)}$	$0.9678^{(5)}$	$0.9885^{(3)}$	$0.9973^{(3)}$
two-body	QID-AWM (no mech.)	$0.9475^{(5)}$	$0.9713^{(5)}$	$0.9880^{(5)}$	$0.9967^{(3)}$
two-body	QID-AWM (scrambled mech.)	$0.9431^{(5)}$	$0.9713^{(5)}$	$0.9878^{(5)}$	–
two-body	B1+KHead	$0.9363^{(8)}$	$0.9643^{(8)}$	$0.9870^{(3)}$	–
two-body	QID-AWM	$0.9437^{(5)}$	$0.9743^{(5)}$	$0.9903^{(5)}$	$0.9970^{(5)}$
spring (cont.)	plain	$0.9025^{(5)}$	$0.9556^{(5)}$	$0.9709^{(5)}$	$0.9876^{(5)}$
spring (cont.)	+MechAux	$0.8846^{(5)}$	$0.9523^{(5)}$	$0.9707^{(3)}$	$0.9877^{(3)}$
spring (cont.)	QID-AWM (no mech.)	$0.9262^{(5)}$	$0.9642^{(5)}$	$0.9739^{(5)}$	$0.9876^{(3)}$
spring (cont.)	QID-AWM (scrambled mech.)	$0.9298^{(5)}$	$0.9637^{(5)}$	$0.9748^{(5)}$	–
spring (cont.)	B1+KHead	$0.9031^{(5)}$	$0.9563^{(5)}$	–	–
spring (cont.)	QID-AWM	$0.9115^{(5)}$	$0.9628^{(5)}$	$0.9745^{(5)}$	$0.9878^{(5)}$
SCM dose	plain	$0.9120^{(5)}$	$0.9515^{(5)}$	$0.9810^{(5)}$	$0.9943^{(5)}$
SCM dose	+MechAux	$0.8985^{(5)}$	$0.9572^{(5)}$	$0.9847^{(3)}$	$0.9953^{(3)}$
SCM dose	QID-AWM (no mech.)	$0.9398^{(5)}$	$0.9635^{(5)}$	$0.9840^{(5)}$	$0.9928^{(3)}$
SCM dose	QID-AWM (scrambled mech.)	$0.9375^{(5)}$	$0.9635^{(5)}$	$0.9844^{(5)}$	–
SCM dose	B1+KHead	$0.9101^{(5)}$	$0.9517^{(5)}$	$0.9814^{(3)}$	–
SCM dose	QID-AWM	$0.9362^{(5)}$	$0.9729^{(5)}$	$0.9884^{(5)}$	$0.9949^{(5)}$

Table 22: Model × loss matrix: QID-AWM minus B1+ID identifiability AUROC (IID) under four scalar-loss recipes. Superscripts give seeds; ≡ marks a 95% CI inside ± 0.002, ≠ a CI excluding zero.

env	raw-unit	tolerance	relative	sym. log
ramp	$+0.0094^{(5)}$ ≠	$-0.0001^{(5)}$ ≡	$+0.0052^{(5)}$ ≠	$+0.0055^{(5)}$ ≠
spring	$-0.0001^{(5)}$ ≡	$-0.0004^{(5)}$ ≡	$+0.0002^{(5)}$ ≡	$-0.0000^{(5)}$ ≡
vis. ramp	$+0.0088^{(5)}$ ≠	$+0.0008^{(10)}$ ≡	$+0.0104^{(5)}$ ≠	$+0.0100^{(5)}$ ≠
vis. spring	$+0.0097^{(5)}$ ≠	$+0.0051^{(5)}$ ≠	$+0.0090^{(5)}$ ≠	$+0.0099^{(5)}$ ≠
pendulum	$-0.0002^{(5)}$ ≡	$-0.0003^{(5)}$ ≡	$-0.0002^{(5)}$ ≡	$-0.0001^{(5)}$ ≡
two-body	$-0.0001^{(5)}$ ≡	$+0.0000^{(5)}$ ≡	$+0.0001^{(5)}$ ≡	$-0.0002^{(5)}$ ≡

Table 23: Learning-rate robustness of the loss result (tolerance-scaled loss, IID identifiability AUROC, QID-AWM / B1+ID with the paired difference and verdict; five seeds, ten at the default rate on the visual ramp). The equivalence does not depend on the optimiser setting carried over from the raw-unit loss.

env	lr 10^{-4}	lr 3×10^{-4} (default)	lr 10^{-3}
ramp	0.9789 / 0.9786 ($+0.0003^{(5)}$ ≡)	0.9806 / 0.9807 ($-0.0001^{(5)}$ ≡)	0.9814 / 0.9816 ($-0.0002^{(5)}$ ≡)
vis. ramp	0.9771 / 0.9771 ($-0.0000^{(5)}$ ≡)	0.9776 / 0.9768 ($+0.0008^{(10)}$ ≡)	0.9781 / 0.9781 ($-0.0001^{(5)}$ ≡)

Table 24: Bank misspecification (tolerance-scaled loss, five seeds, 800 episodes per system). Test episodes are generated from one law outside each training bank; labels are computed against the augmented bank (reality). "silent": the training bank calls the query identifiable but reality does not, or the bank's answer is wrong; "detect.": the bank has no compatible hypothesis; "honest": bank and reality agree. p(id) is the model's mean predicted identifiability; "called id." the fraction it calls identifiable at 0.5.

env	model	seeds	silent	detect.	AUROC (true)	p(id) silent	called id.	cf-acc silent	p(id) honest	cf-acc honest
ramp	QID-AWM	5	0.02	0.00	0.973	0.35	0.27	0.37	0.90	0.91
ramp	B1+ID	5	0.02	0.00	0.972	0.40	0.33	0.39	0.92	0.91
spring	QID-AWM	5	0.04	0.00	0.978	0.34	0.31	0.50	0.93	0.97
spring	B1+ID	5	0.04	0.00	0.979	0.39	0.35	0.50	0.94	0.97
pendulum	QID-AWM	5	0.01	0.00	0.989	0.40	0.36	0.81	0.94	0.98
pendulum	B1+ID	5	0.01	0.00	0.989	0.45	0.39	0.86	0.95	0.98
two-body	QID-AWM	5	0.01	0.00	0.997	0.92	0.92	1.00	0.96	1.00
two-body	B1+ID	5	0.01	0.00	0.996	0.94	0.94	0.99	0.97	1.00
SCM dose	QID-AWM	5	0.02	0.00	0.994	0.26	0.22	0.73	0.96	0.97
SCM dose	B1+ID	5	0.02	0.00	0.993	0.30	0.28	0.66	0.97	0.98

Table 25: Ablations under the tolerance-scaled loss on four environments, including the two systems added for this paper: change in identifiability AUROC when a component is removed (three seeds). Only supervision and the query itself matter.

removed	spring	vis. ramp	pendulum	two-body
early query fusion	$-0.005^{(3)}$	$-0.001^{(3)}$	$-0.003^{(3)}$	$-0.003^{(3)}$
mechanism posterior	$+0.000^{(3)}$	$+0.001^{(3)}$	$-0.000^{(3)}$	$+0.000^{(3)}$
mechanism supervision	$-0.000^{(3)}$	$-0.000^{(3)}$	$-0.000^{(3)}$	$-0.000^{(3)}$
dispersion supervision	$+0.000^{(3)}$	$+0.001^{(3)}$	$-0.000^{(3)}$	$+0.000^{(3)}$
cost penalty	$+0.000^{(3)}$	$+0.000^{(3)}$	$-0.000^{(3)}$	$+0.000^{(3)}$
evidence scoring	$-0.001^{(3)}$	$+0.001^{(3)}$	$-0.002^{(3)}$	$+0.000^{(3)}$
calibration	$+0.000^{(3)}$	$-0.001^{(3)}$	$-0.000^{(3)}$	$+0.000^{(3)}$
identifiability-specific loss	$-0.003^{(3)}$	$-0.001^{(3)}$	$-0.004^{(3)}$	$-0.002^{(3)}$
identifiability supervision	$-0.350^{(3)}$	$-0.402^{(3)}$	$-0.469^{(3)}$	$-0.584^{(3)}$
the query itself	$-0.292^{(3)}$	$-0.338^{(3)}$	$-0.333^{(3)}$	$-0.265^{(3)}$

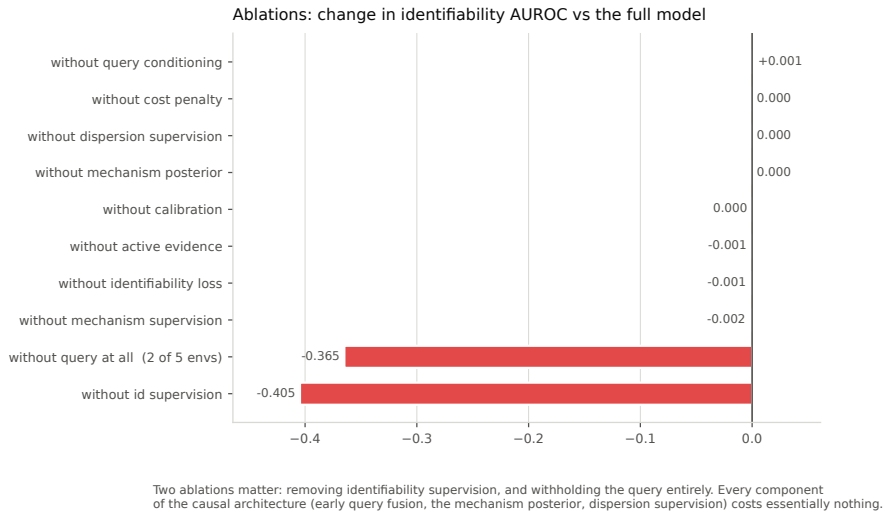

Figure 4: Ablation deltas in identifiability AUROC against the full model, mean over environments and seeds (raw-unit loss, the five original environments; Table 25 repeats the ablations under the tolerance-scaled loss on four environments with the same outcome). Only supervision matters; withholding the query entirely is the other exception.

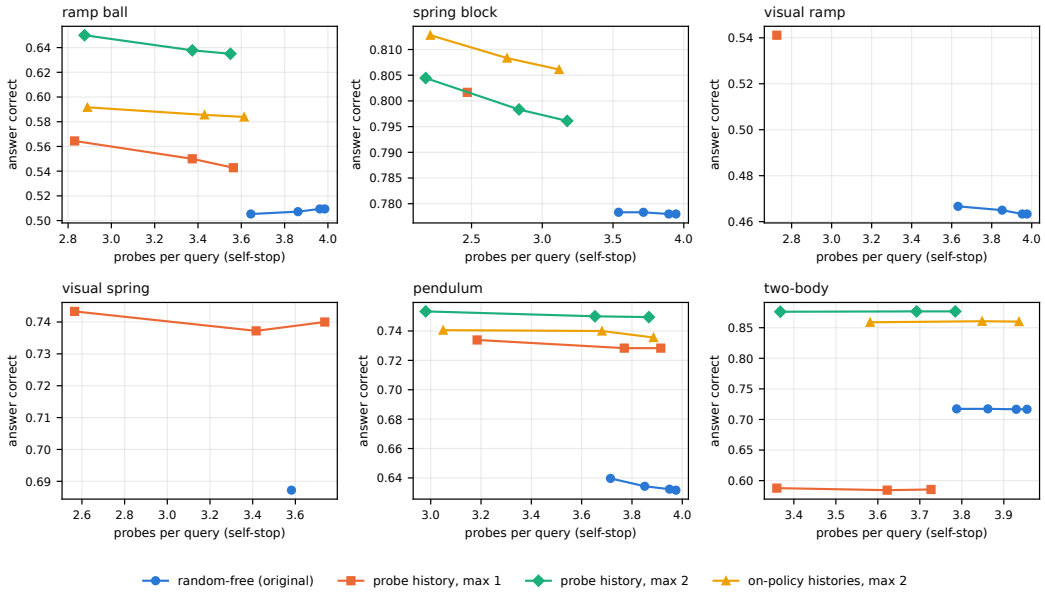

Figure 5: Probes spent against answer accuracy under the agent's own stopping rule (thresholds 0.5–0.97 along each curve), for the tolerance-loss model and the probe-history and on-policy-history variants. Random appended probes buy fewer probes at similar accuracy; on-policy histories buy both.

Table 26: The agent's own stopping rule. Threshold sweeps change probes and resolution but not answer accuracy; probe-history training changes what the agent can do after probing.

env	training	stop thr.	seeds	probes	resolved	oracle-stop resolved	answer correct
ramp	tolerance-loss	0.5	5	3.64	0.930	0.997	0.505
ramp	tolerance-loss	0.7	3	3.86	0.976	0.997	0.507
ramp	tolerance-loss	0.9	3	3.96	0.991	0.997	0.509
ramp	tolerance-loss	0.97	3	3.99	0.994	0.997	0.509
ramp	+ probe-history (max 1)	0.5	3	2.83	0.886	0.995	0.564
ramp	+ probe-history (max 1)	0.9	3	3.37	0.978	0.995	0.550
ramp	+ probe-history (max 1)	0.97	3	3.56	0.987	0.995	0.543
ramp	+ probe-history (max 2)	0.5	3	2.88	0.923	0.997	0.650
ramp	+ probe-history (max 2)	0.9	3	3.37	0.988	0.997	0.638
ramp	+ probe-history (max 2)	0.97	3	3.55	0.996	0.997	0.635
ramp	+ on-policy histories (max 2)	0.5	3	2.89	0.913	0.996	0.592
ramp	+ on-policy histories (max 2)	0.9	3	3.43	0.988	0.996	0.586
ramp	+ on-policy histories (max 2)	0.97	3	3.61	0.994	0.996	0.584
spring	tolerance-loss	0.5	5	3.54	0.935	0.989	0.778
spring	tolerance-loss	0.7	5	3.71	0.954	0.989	0.778
spring	tolerance-loss	0.9	5	3.89	0.981	0.989	0.778
spring	tolerance-loss	0.97	5	3.94	0.986	0.989	0.778
spring	+ probe-history (max 3)	0.5	3	2.09	0.724	0.976	0.798
spring	+ probe-history (max 3)	0.7	3	2.37	0.772	0.976	0.798
spring	+ probe-history (max 3)	0.9	3	2.75	0.843	0.976	0.798
spring	+ probe-history (max 3)	0.97	3	3.09	0.906	0.976	0.794
spring	+ probe-history (max 1)	0.5	3	2.47	0.827	0.989	0.802
spring	+ probe-history (max 2)	0.5	3	2.18	0.772	0.983	0.804
spring	+ probe-history (max 2)	0.9	3	2.84	0.899	0.983	0.798
spring	+ probe-history (max 2)	0.97	3	3.18	0.941	0.983	0.796
spring	+ on-policy histories (max 2)	0.5	3	2.21	0.889	0.984	0.813
spring	+ on-policy histories (max 2)	0.9	3	2.75	0.958	0.984	0.808
spring	+ on-policy histories (max 2)	0.97	3	3.12	0.971	0.984	0.806
vis. ramp	tolerance-loss	0.5	3	3.63	0.916	0.997	0.467
vis. ramp	tolerance-loss	0.7	3	3.85	0.968	0.997	0.465
vis. ramp	tolerance-loss	0.9	3	3.95	0.990	0.997	0.463
vis. ramp	tolerance-loss	0.97	3	3.98	0.993	0.997	0.463
vis. ramp	+ probe-history (max 3)	0.5	3	2.65	0.854	0.994	0.554
vis. ramp	+ probe-history (max 3)	0.7	3	2.83	0.893	0.994	0.554
vis. ramp	+ probe-history (max 3)	0.9	3	3.07	0.934	0.994	0.553
vis. ramp	+ probe-history (max 3)	0.97	3	3.30	0.962	0.994	0.551
vis. ramp	+ probe-history (max 1)	0.5	3	2.73	0.856	0.994	0.541
vis. spring	tolerance-loss	0.5	3	3.58	0.909	0.986	0.687
vis. spring	+ probe-history (max 1)	0.5	3	2.57	0.788	0.977	0.743
vis. spring	+ probe-history (max 1)	0.9	3	3.42	0.929	0.977	0.737
vis. spring	+ probe-history (max 1)	0.97	3	3.74	0.956	0.977	0.740
pendulum	tolerance-loss	0.5	5	3.71	0.940	0.977	0.640
pendulum	tolerance-loss	0.7	5	3.85	0.961	0.977	0.634
pendulum	tolerance-loss	0.9	5	3.95	0.975	0.977	0.632
pendulum	tolerance-loss	0.97	5	3.97	0.977	0.977	0.632
pendulum	+ probe-history (max 1)	0.5	3	3.18	0.882	0.983	0.734
pendulum	+ probe-history (max 1)	0.9	3	3.77	0.962	0.983	0.728
pendulum	+ probe-history (max 1)	0.97	3	3.91	0.980	0.983	0.728
pendulum	+ probe-history (max 2)	0.5	3	2.98	0.866	0.978	0.753
pendulum	+ probe-history (max 2)	0.9	3	3.65	0.956	0.978	0.750
pendulum	+ probe-history (max 2)	0.97	3	3.87	0.975	0.978	0.749
pendulum	+ on-policy histories (max 2)	0.5	3	3.05	0.892	0.976	0.741
pendulum	+ on-policy histories (max 2)	0.9	3	3.68	0.965	0.976	0.740
pendulum	+ on-policy histories (max 2)	0.97	3	3.89	0.972	0.976	0.736
two-body	tolerance-loss	0.5	5	3.79	0.890	0.911	0.717
two-body	tolerance-loss	0.7	5	3.86	0.900	0.911	0.717
two-body	tolerance-loss	0.9	5	3.93	0.908	0.911	0.717
two-body	tolerance-loss	0.97	5	3.96	0.909	0.911	0.717
two-body	+ probe-history (max 1)	0.5	3	3.36	0.858	0.904	0.588
two-body	+ probe-history (max 1)	0.9	3	3.62	0.891	0.904	0.584
two-body	+ probe-history (max 1)	0.97	3	3.73	0.896	0.904	0.586
two-body	+ probe-history (max 2)	0.5	3	3.37	0.845	0.898	0.876
two-body	+ probe-history (max 2)	0.9	3	3.69	0.886	0.898	0.877
two-body	+ probe-history (max 2)	0.97	3	3.78	0.893	0.898	0.877
two-body	+ on-policy histories (max 2)	0.5	3	3.58	0.859	0.902	0.859
two-body	+ on-policy histories (max 2)	0.9	3	3.85	0.892	0.902	0.861
two-body	+ on-policy histories (max 2)	0.97	3	3.94	0.898	0.902	0.860

Table 27: The null is robust to capacity, training length and posterior size (tolerance-scaled loss, IID AUROC). Particle-count cells list AUROC for $K = 2, 4, 6, 12, 24$.

env	setting	seeds	QID-AWM	B1+ID	Δ
ramp	d=128, 2 layers	3	0.9795	0.9793	+0.0003
ramp	d=256, 4 layers (default)	5	0.9806	0.9807	-0.0001
ramp	d=512, 6 layers	3	0.9810	0.9822	-0.0011
ramp	d=1024, 6 layers	3	0.9810	0.9812	-0.0002
ramp	default, 120 epochs	3	0.9836	0.9828	+0.0007
ramp	particles $K \in \{2, 4, 6, 12, 24\}$, unsup.	3	0.9805 / 0.9809 / 0.9809 / 0.9803 / 0.9804		
spring	d=128, 2 layers	3	0.9820	0.9805	+0.0015
spring	d=256, 4 layers (default)	5	0.9836	0.9840	-0.0004
spring	d=512, 6 layers	3	0.9839	0.9831	+0.0008
spring	d=1024, 6 layers	3	0.9842	0.9844	-0.0003
spring	default, 120 epochs	3	0.9824	0.9822	+0.0002
spring	particles $K \in \{2, 4, 6, 12, 24\}$, unsup.	3	0.9838 / 0.9835 / 0.9836 / 0.9838 / 0.9838		

Table 28: Sequential evidence acquisition, full oracle-policy columns (QID-AWM, tolerance-scaled loss, ambiguous IID queries, ground-truth stopping): probes to resolution under the exact query-specific oracle, generic information gain, RAED-style structural gain, random, and the learned agent; the last column is the paired 95% interval on the oracle−generic probe difference.

env	seeds	oracle	generic	structural	random	learned	vs generic	vs oracle	oracle − generic, 95% CI
collision	5	1.76	2.97	3.05	3.08	1.78	-40.0%	+1.0%	[-1.28, -1.07]
ramp	5	1.03	1.14	1.21	1.40	1.05	-8.4%	+1.3%	[-0.13, -0.05]
spring	5	1.06	1.32	1.26	1.48	1.10	-16.9%	+3.2%	[-0.27, -0.17]
vis. collision	3	1.74	2.97	3.07	3.14	1.76	-40.6%	+1.6%	[-1.31, -1.10]
vis. ramp	3	1.03	1.13	1.19	1.40	1.06	-6.6%	+2.8%	[-0.11, -0.03]
vis. spring	3	1.06	1.33	1.24	1.50	1.11	-16.6%	+4.9%	[-0.27, -0.17]
pendulum	5	1.31	1.34	1.60	2.06	1.40	+4.6%	+6.3%	[+0.01, +0.14]
two-body	5	1.26	1.79	1.79	1.93	1.38	-22.7%	+9.9%	[-0.47, -0.31]
spring (cont.)	3	1.07	1.25	1.55	1.66	1.09	-13.0%	+2.2%	[-0.20, -0.11]
SCM dose	3	1.11	1.21	1.34	1.57	1.19	-1.6%	+7.3%	[-0.05, +0.04]

Table 29: Learned identifiability head against the posterior's own dispersion threshold (tolerance-scaled loss, IID identifiability AUROC, mean±std over seeds). The threshold scores each query by the model's predicted answer dispersion across mechanism hypotheses, with no learned head, rank-normalised per target type, which favours it.

env	seeds	learned head	D_q threshold	head − threshold
collision	5	1.000±0.000	0.997±0.001	+0.002
ramp	5	0.981±0.000	0.690±0.007	+0.290
spring	5	0.984±0.000	0.661±0.018	+0.323
vis. collision	5	1.000±0.000	0.997±0.000	+0.003
vis. ramp	10	0.978±0.000	0.693±0.009	+0.285
vis. spring	5	0.977±0.001	0.620±0.021	+0.357
pendulum	5	0.992±0.000	0.754±0.020	+0.238
two-body	5	0.997±0.000	0.769±0.005	+0.228
spring (cont.)	5	0.988±0.000	0.675±0.012	+0.313
SCM dose	5	0.995±0.000	0.583±0.021	+0.412

Table 30: Calibration and selective prediction under the tolerance-scaled loss (IID split, five seeds): expected calibration error / area under the risk–coverage curve for the confidence-based baselines B1–B3, the plain control B1+ID and QID-AWM.

env	seeds	B1	B2 ensemble	B3 MC dropout	B1+ID	QID-AWM
collision	5	0.282 / 0.008	0.332 / 0.010	0.327 / 0.010	0.005 / 0.011	0.004 / 0.011
ramp	5	0.303 / 0.053	0.294 / 0.052	0.312 / 0.056	0.010 / 0.034	0.025 / 0.035
spring	5	0.198 / 0.033	0.331 / 0.027	0.353 / 0.027	0.012 / 0.019	0.018 / 0.019
vis. collision	5	0.286 / 0.009	0.325 / 0.010	0.320 / 0.010	0.006 / 0.011	0.005 / 0.011
vis. ramp	5	0.290 / 0.067	0.289 / 0.067	0.308 / 0.073	0.026 / 0.041	0.029 / 0.044
vis. spring	5	0.212 / 0.038	0.326 / 0.033	0.347 / 0.035	0.059 / 0.022	0.023 / 0.021
pendulum	5	0.283 / 0.038	0.306 / 0.038	0.303 / 0.038	0.007 / 0.020	0.020 / 0.020
two-body	5	0.285 / 0.007	0.273 / 0.007	0.286 / 0.007	0.008 / 0.005	0.015 / 0.005
spring (cont.)	5	0.201 / 0.023	0.339 / 0.021	0.356 / 0.020	0.016 / 0.014	0.016 / 0.014
SCM dose	5	0.255 / 0.087	0.267 / 0.088	0.258 / 0.092	0.007 / 0.024	0.021 / 0.033

Table 31: Split-conformal prediction at $\alpha = 0.10$ on the tolerance-scaled checkpoints, five seeds, all environments. Scalar width ratio is non-identifiable over identifiable interval width; coverage is reported overall and on the non-identifiable subgroup.

env	split	seeds	width ratio	coverage (all)	coverage (non-id.)
ramp	IID	5	2.04 ± 0.13	0.909	0.869
ramp	mechanism-OOD	5	1.69 ± 0.16	0.678	0.617
ramp	composition-OOD	5	1.43 ± 0.02	0.411	0.410
spring	IID	5	1.18 ± 0.07	0.892	0.849
spring	mechanism-OOD	5	1.13 ± 0.05	0.921	0.898
spring	composition-OOD	5	1.12 ± 0.05	0.842	0.781
vis. ramp	IID	5	2.02 ± 0.08	0.912	0.871
vis. ramp	mechanism-OOD	5	1.65 ± 0.22	0.691	0.629
vis. ramp	composition-OOD	5	1.51 ± 0.05	0.429	0.429
vis. spring	IID	5	1.16 ± 0.11	0.893	0.855
vis. spring	mechanism-OOD	5	1.10 ± 0.09	0.908	0.883
vis. spring	composition-OOD	5	1.12 ± 0.08	0.834	0.773
pendulum	IID	5	1.27 ± 0.08	0.900	0.824
pendulum	mechanism-OOD	5	1.10 ± 0.06	0.942	0.896
pendulum	composition-OOD	5	0.87 ± 0.06	0.806	0.732
two-body	IID	5	3.96 ± 0.50	0.900	0.925
two-body	mechanism-OOD	5	3.83 ± 0.57	0.856	0.834
two-body	composition-OOD	5	4.05 ± 0.45	0.802	0.821
spring (cont.)	IID	5	1.18 ± 0.14	0.890	0.857
spring (cont.)	mechanism-OOD	5	1.03 ± 0.02	0.651	0.564
spring (cont.)	composition-OOD	5	1.10 ± 0.07	0.578	0.480
SCM dose	IID	5	0.93 ± 0.08	0.882	0.774
SCM dose	mechanism-OOD	5	0.91 ± 0.08	0.866	0.758
SCM dose	composition-OOD	5	2.60 ± 0.34	0.819	0.759

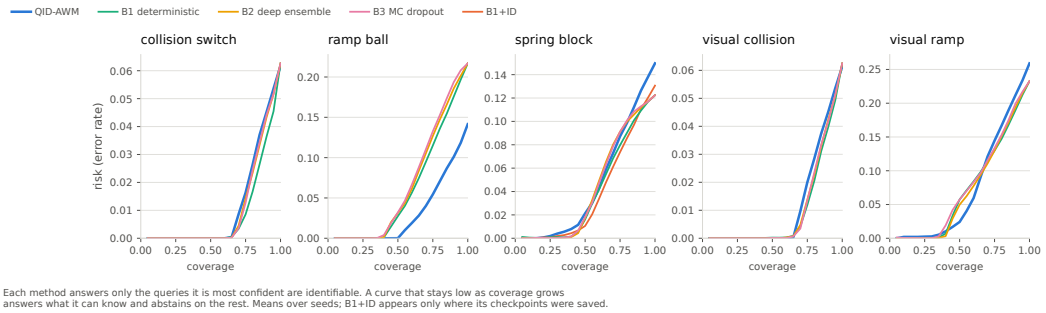

Figure 6: Risk–coverage for selective prediction on the IID split, means over seeds, scored with each target's own answer tolerance.

C RAW-UNIT-LOSS RESULTS (SUPERSEDED; RETAINED FOR THE AUDIT TRAIL)

Everything in this section was computed under the raw-unit counterfactual loss that Section 4 shows to handicap the plain control. It is retained because the paper's argument depends on the reader being able to see the retracted comparison. The main text cites it only for the retracted comparison itself; the tolerance-scaled counterparts are Tables 2, 18, 19, 30, 31 and 29. Figures 7 and 8 are the raw-loss baseline comparison and the oracle acquisition criteria; a hyphen in a table marks a baseline not run on that split.

Table 32: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Main IID results, identifiability AUROC (mean±std over seeds; "–" where a confidence-based baseline was not run).

env	QID-AWM	B1+ID	B1	B2	B3	B4	B0
collision	1.000±0.000	1.000±0.000	0.998±0.000	0.996±0.000	0.996±0.000	0.996±0.000	0.741±0.000
pendulum	0.993±0.000	0.993±0.000	–	–	–	–	0.693±0.000
ramp	0.993±0.003	0.984±0.001	0.907±0.008	0.902±0.006	0.889±0.019	0.902±0.006	0.654±0.000
SCM dose	0.996±0.000	0.996±0.000	–	–	–	–	0.792±0.000
spring	0.983±0.000	0.983±0.000	0.896±0.007	0.850±0.015	0.856±0.008	0.850±0.015	0.565±0.000
spring (cont.)	0.988±0.000	0.987±0.000	–	–	–	–	0.601±0.000
two-body	0.997±0.000	0.997±0.000	–	–	–	–	0.620±0.000
vis. collision	1.000±0.000	1.000±0.000	0.996±0.000	0.995±0.000	0.994±0.000	0.995±0.000	0.741±0.000
vis. ramp	0.977±0.001	0.968±0.001	0.899±0.010	0.892±0.005	0.886±0.009	0.892±0.005	0.657±0.000
vis. spring	0.976±0.000	0.967±0.001	–	–	–	–	0.565±0.000

Table 33: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Main IID results, counterfactual accuracy (mean±std over seeds).

env	QID-AWM	B1+ID	B1	B2	B3	B4	B0
collision	0.938±0.000	0.938±0.000	0.937±0.001	0.937±0.001	0.937±0.001	0.937±0.001	–
pendulum	0.786±0.003	0.807±0.008	–	–	–	–	–
ramp	0.859±0.027	0.782±0.003	0.783±0.003	0.783±0.003	0.783±0.003	0.783±0.003	–
SCM dose	0.737±0.024	0.821±0.006	–	–	–	–	–
spring	0.850±0.011	0.869±0.002	0.878±0.003	0.878±0.003	0.878±0.003	0.878±0.003	–
spring (cont.)	0.880±0.003	0.887±0.002	–	–	–	–	–
two-body	0.865±0.013	0.914±0.003	–	–	–	–	–
vis. collision	0.939±0.000	0.938±0.001	0.936±0.001	0.936±0.001	0.936±0.001	0.936±0.001	–
vis. ramp	0.741±0.019	0.767±0.006	0.764±0.001	0.764±0.001	0.764±0.001	0.764±0.001	–
vis. spring	0.762±0.024	0.832±0.005	–	–	–	–	–

Table 34: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Main IID results, area under the risk–coverage curve, lower is better (mean±std over seeds).

env	QID-AWM	B1+ID	B1	B2	B3	B4	B0
collision	0.011±0.000	0.011±0.000	0.008±0.000	0.010±0.000	0.010±0.000	0.010±0.000	0.025±0.000
pendulum	0.094±0.011	0.068±0.004	–	–	–	–	0.183±0.007
ramp	0.030±0.008	0.053±0.001	0.060±0.003	0.064±0.003	0.067±0.005	0.064±0.003	0.082±0.017
SCM dose	0.117±0.012	0.050±0.004	–	–	–	–	0.239±0.024
spring	0.045±0.005	0.037±0.001	0.041±0.002	0.043±0.002	0.044±0.003	0.043±0.002	0.137±0.017
spring (cont.)	0.028±0.001	0.025±0.001	–	–	–	–	0.100±0.003
two-body	0.033±0.004	0.020±0.001	–	–	–	–	0.094±0.013
vis. collision	0.011±0.000	0.011±0.000	0.009±0.000	0.010±0.000	0.010±0.000	0.010±0.000	0.025±0.000
vis. ramp	0.073±0.009	0.058±0.002	0.072±0.004	0.071±0.002	0.073±0.003	0.071±0.002	0.191±0.028
vis. spring	0.077±0.009	0.055±0.002	–	–	–	–	0.229±0.029

Table 35: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Calibration and selective prediction

Env	Method	ECE	Brier	AURC	acc@50	acc@70	acc@90
collision	**QID-AWM**	0.003±0.000	0.006±0.000	0.011±0.000	1.000±0.000	0.991±0.001	0.955±0.002
collision	B1 deterministic WM	0.282±0.000	0.238±0.000	0.008±0.000	1.000±0.000	0.997±0.000	0.964±0.001
collision	B2 deep ensemble	0.332±0.000	0.323±0.001	0.010±0.000	1.000±0.000	0.996±0.001	0.957±0.001
collision	B3 MC dropout	0.327±0.001	0.314±0.001	0.010±0.000	1.000±0.000	0.996±0.001	0.957±0.001
collision	B4 ensemble disagreement	0.098±0.010	0.659±0.000	0.010±0.000	1.000±0.000	0.996±0.001	0.957±0.001
pendulum	**QID-AWM**	0.021±0.001	0.033±0.001	0.094±0.011	0.943±0.005	0.891±0.003	0.810±0.004
pendulum	B1 deterministic WM	–	–	–	–	–	–
pendulum	B2 deep ensemble	–	–	–	–	–	–
pendulum	B3 MC dropout	–	–	–	–	–	–
pendulum	B4 ensemble disagreement	–	–	–	–	–	–
ramp	**QID-AWM**	0.013±0.005	0.032±0.006	0.030±0.008	1.000±0.000	0.960±0.014	0.899±0.028
ramp	B1 deterministic WM	0.259±0.027	0.238±0.023	0.060±0.003	0.973±0.012	0.905±0.004	0.824±0.004
ramp	B2 deep ensemble	0.277±0.019	0.257±0.018	0.064±0.003	0.969±0.006	0.895±0.011	0.814±0.005
ramp	B3 MC dropout	0.310±0.005	0.297±0.004	0.067±0.005	0.970±0.012	0.889±0.011	0.809±0.006
ramp	B4 ensemble disagreement	0.055±0.020	0.647±0.001	0.064±0.003	0.969±0.006	0.895±0.011	0.814±0.005
SCM dose	**QID-AWM**	0.014±0.001	0.024±0.001	0.117±0.012	0.877±0.021	0.867±0.021	0.784±0.018
SCM dose	B1 deterministic WM	–	–	–	–	–	–
SCM dose	B2 deep ensemble	–	–	–	–	–	–
SCM dose	B3 MC dropout	–	–	–	–	–	–
SCM dose	B4 ensemble disagreement	–	–	–	–	–	–
spring	**QID-AWM**	0.021±0.002	0.052±0.001	0.045±0.005	0.979±0.006	0.927±0.008	0.874±0.011
spring	B1 deterministic WM	0.269±0.007	0.246±0.007	0.041±0.002	0.982±0.002	0.923±0.006	0.892±0.003
spring	B2 deep ensemble	0.275±0.002	0.265±0.003	0.043±0.002	0.981±0.002	0.916±0.006	0.886±0.004
spring	B3 MC dropout	0.321±0.004	0.311±0.004	0.044±0.003	0.979±0.003	0.915±0.008	0.888±0.003
spring	B4 ensemble disagreement	0.212±0.020	0.621±0.002	0.043±0.002	0.981±0.002	0.916±0.006	0.886±0.004
spring (cont.)	**QID-AWM**	0.017±0.002	0.040±0.001	0.028±0.001	0.998±0.002	0.961±0.003	0.910±0.002
spring (cont.)	B1 deterministic WM	–	–	–	–	–	–
spring (cont.)	B2 deep ensemble	–	–	–	–	–	–
spring (cont.)	B3 MC dropout	–	–	–	–	–	–
spring (cont.)	B4 ensemble disagreement	–	–	–	–	–	–
two-body	**QID-AWM**	0.019±0.001	0.024±0.000	0.033±0.004	0.987±0.006	0.961±0.007	0.891±0.009
two-body	B1 deterministic WM	–	–	–	–	–	–
two-body	B2 deep ensemble	–	–	–	–	–	–
two-body	B3 MC dropout	–	–	–	–	–	–
two-body	B4 ensemble disagreement	–	–	–	–	–	–
vis. collision	**QID-AWM**	0.004±0.001	0.006±0.000	0.011±0.000	1.000±0.000	0.990±0.002	0.955±0.002
vis. collision	B1 deterministic WM	0.285±0.000	0.244±0.001	0.009±0.000	1.000±0.000	0.995±0.001	0.961±0.001
vis. collision	B2 deep ensemble	0.325±0.001	0.311±0.002	0.010±0.000	1.000±0.000	0.996±0.000	0.956±0.001
vis. collision	B3 MC dropout	0.321±0.001	0.304±0.002	0.010±0.000	1.000±0.000	0.996±0.001	0.958±0.000
vis. collision	B4 ensemble disagreement	0.152±0.026	0.659±0.000	0.010±0.000	1.000±0.000	0.996±0.000	0.956±0.001
vis. ramp	**QID-AWM**	0.031±0.010	0.062±0.003	0.073±0.009	0.976±0.015	0.879±0.018	0.788±0.017
vis. ramp	B1 deterministic WM	0.295±0.016	0.277±0.021	0.072±0.004	0.950±0.009	0.886±0.008	0.808±0.005
vis. ramp	B2 deep ensemble	0.251±0.004	0.234±0.004	0.071±0.002	0.951±0.004	0.888±0.006	0.806±0.003
vis. ramp	B3 MC dropout	0.309±0.003	0.297±0.005	0.073±0.003	0.948±0.009	0.885±0.009	0.801±0.004
vis. ramp	B4 ensemble disagreement	0.185±0.007	0.652±0.000	0.071±0.002	0.951±0.004	0.888±0.006	0.806±0.003
vis. spring	**QID-AWM**	0.028±0.006	0.062±0.001	0.077±0.009	0.954±0.012	0.873±0.015	0.794±0.022
vis. spring	B1 deterministic WM	–	–	–	–	–	–
vis. spring	B2 deep ensemble	–	–	–	–	–	–
vis. spring	B3 MC dropout	–	–	–	–	–	–
vis. spring	B4 ensemble disagreement	–	–	–	–	–	–

Table 36: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Learned identifiability head vs raw dispersion threshold. Identifiability AUROC, mean±std over seeds. The raw threshold scores each query by the model's own answer dispersion across mechanism hypotheses, with no learned identifiability head, rank-normalised per target type (which favours it).

Env	Split	learned head	raw D_q threshold	head - raw
collision	IID	1.000±0.000	0.998±0.000	+0.0018
collision	Mechanism-OOD	1.000±0.000	0.999±0.000	+0.0011
pendulum	IID	0.993±0.000	0.728±0.029	+0.2644
pendulum	Mechanism-OOD	0.992±0.001	0.738±0.023	+0.2537
ramp	IID	0.993±0.003	0.637±0.039	+0.3557
ramp	Mechanism-OOD	0.990±0.002	0.617±0.022	+0.3729
SCM dose	IID	0.996±0.000	0.609±0.046	+0.3866
SCM dose	Mechanism-OOD	0.993±0.001	0.579±0.049	+0.4144
spring	IID	0.983±0.000	0.564±0.032	+0.4185
spring	Mechanism-OOD	0.986±0.001	0.605±0.028	+0.3813
spring (cont.)	IID	0.988±0.000	0.620±0.028	+0.3671
spring (cont.)	Mechanism-OOD	0.967±0.001	0.574±0.015	+0.3926
two-body	IID	0.997±0.000	0.719±0.016	+0.2780
two-body	Mechanism-OOD	0.996±0.000	0.698±0.019	+0.2979
vis. collision	IID	1.000±0.000	0.997±0.000	+0.0027
vis. collision	Mechanism-OOD	1.000±0.000	0.999±0.000	+0.0008
vis. ramp	IID	0.977±0.001	0.676±0.020	+0.3014
vis. ramp	Mechanism-OOD	0.978±0.001	0.611±0.018	+0.3671
vis. spring	IID	0.976±0.000	0.582±0.057	+0.3943
vis. spring	Mechanism-OOD	0.980±0.001	0.564±0.030	+0.4163

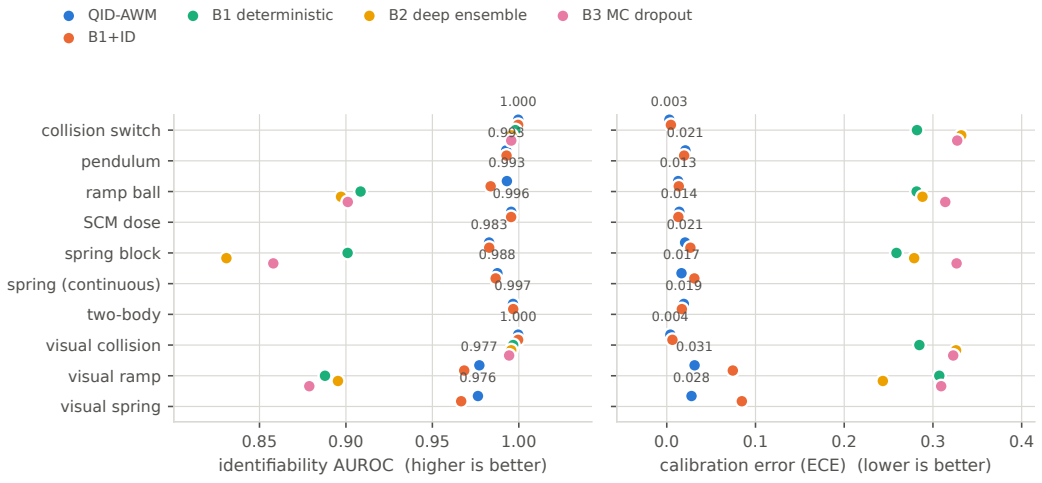

Figure 7: Identifiability AUROC and calibration error for all methods (raw-unit loss).

Table 37: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). OOD results. Identifiability AUROC, mean±std over seeds.

Env	Split	QID-AWM	B1 deterministic WM	B2 deep ensemble	B3 MC dropout	B4 ensemble disagreement
collision	IID	1.000±0.000	0.998±0.000	0.996±0.000	0.996±0.000	0.996±0.000
collision	Parameter-OOD	1.000±0.000	0.993±0.001	0.990±0.002	0.987±0.004	0.990±0.002
collision	Mechanism-OOD	1.000±0.000	0.999±0.000	0.997±0.000	0.997±0.000	0.997±0.000
collision	Composition-OOD	1.000±0.000	0.997±0.000	0.995±0.000	0.994±0.001	0.995±0.000
pendulum	IID	0.993±0.000	–	–	–	–
pendulum	Parameter-OOD	0.992±0.001	–	–	–	–
pendulum	Mechanism-OOD	0.992±0.001	–	–	–	–
pendulum	Composition-OOD	0.989±0.001	–	–	–	–
ramp	IID	0.993±0.003	0.907±0.008	0.902±0.006	0.889±0.019	0.902±0.006
ramp	Parameter-OOD	0.987±0.004	0.890±0.016	0.877±0.009	0.868±0.015	0.877±0.009
ramp	Mechanism-OOD	0.990±0.002	0.914±0.002	0.926±0.004	0.896±0.015	0.926±0.004
ramp	Composition-OOD	0.974±0.004	0.876±0.010	0.899±0.008	0.855±0.010	0.899±0.008
SCM dose	IID	0.996±0.000	–	–	–	–
SCM dose	Parameter-OOD	0.991±0.001	–	–	–	–
SCM dose	Mechanism-OOD	0.993±0.001	–	–	–	–
SCM dose	Composition-OOD	0.986±0.001	–	–	–	–
spring	IID	0.983±0.000	0.896±0.007	0.850±0.015	0.856±0.008	0.850±0.015
spring	Parameter-OOD	0.979±0.001	0.870±0.011	0.836±0.011	0.838±0.012	0.836±0.012
spring	Mechanism-OOD	0.986±0.001	0.892±0.007	0.857±0.005	0.861±0.005	0.857±0.005
spring	Composition-OOD	0.978±0.001	0.876±0.008	0.849±0.011	0.851±0.006	0.849±0.011
spring (cont.)	IID	0.988±0.000	–	–	–	–
spring (cont.)	Parameter-OOD	0.977±0.001	–	–	–	–
spring (cont.)	Mechanism-OOD	0.967±0.001	–	–	–	–
spring (cont.)	Composition-OOD	0.987±0.000	–	–	–	–
two-body	IID	0.997±0.000	–	–	–	–
two-body	Parameter-OOD	0.992±0.001	–	–	–	–
two-body	Mechanism-OOD	0.996±0.000	–	–	–	–
two-body	Composition-OOD	0.996±0.000	–	–	–	–
vis. collision	IID	1.000±0.000	0.996±0.000	0.995±0.000	0.994±0.000	0.995±0.000
vis. collision	Parameter-OOD	1.000±0.000	0.993±0.001	0.991±0.001	0.987±0.002	0.991±0.001
vis. collision	Mechanism-OOD	1.000±0.000	0.998±0.000	0.997±0.000	0.996±0.000	0.997±0.000
vis. collision	Composition-OOD	1.000±0.000	0.997±0.000	0.996±0.001	0.994±0.001	0.996±0.001
vis. ramp	IID	0.977±0.001	0.899±0.010	0.892±0.005	0.886±0.009	0.892±0.005
vis. ramp	Parameter-OOD	0.930±0.009	0.860±0.026	0.782±0.049	0.820±0.023	0.782±0.049
vis. ramp	Mechanism-OOD	0.978±0.001	0.921±0.007	0.911±0.005	0.897±0.006	0.911±0.005
vis. ramp	Composition-OOD	0.953±0.002	0.895±0.008	0.878±0.004	0.860±0.006	0.878±0.004
vis. spring	IID	0.976±0.000	–	–	–	–
vis. spring	Parameter-OOD	0.971±0.002	–	–	–	–
vis. spring	Mechanism-OOD	0.980±0.001	–	–	–	–
vis. spring	Composition-OOD	0.972±0.001	–	–	–	–

Table 38: Raw-unit training loss (superseded by the tolerance-scaled results in the main text). Visual-OOD (appearance shift, identical physics and labels). Same seeds, controls, mechanisms and labels as the IID split; only the rendered appearance changes (contrast, background gradient, vignette). Mean over seeds.

Env	Split	AUROC	ECE	CF accuracy
vis. collision	IID	0.9997	0.0041	0.9386
vis. collision	Visual-OOD	0.9971	0.0343	0.9273
vis. collision	**shift cost**	**-0.0026**		**-0.0113**
vis. ramp	IID	0.9772	0.0314	0.7410
vis. ramp	Visual-OOD	0.9383	0.0727	0.4590
vis. ramp	**shift cost**	**-0.0388**		**-0.2820**

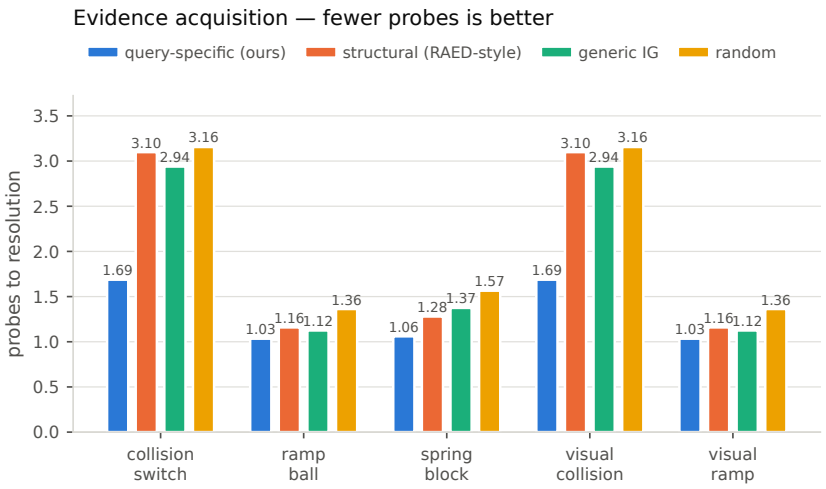

Figure 8: Oracle evidence acquisition by criterion, paired menus (raw-unit loss checkpoints).