Neurosymbolic Artificial Intelligence
XX(X):1–15
©The Author(s) 2026
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

# Supplementary material for "Expressiveness, execution, and lifted recurrent closure in neurosymbolic intervention logic"

**Aliefya Mahalva Ayn**

## S1. Seed-level neural evaluation

This section reports seed-level uncertainty and the quotient-fidelity ablations used to interpret the recurrent experiment. All five seeds use the same train/evaluation sizes as the main text.

*State-disjoint decision probe and architectural control.* For each seed, the 219 exact depth-two conjunction far-class states are stratified by missing-control count and split into 153 probe-training states and 66 held-out probe-test states. The held-out states are also excluded from the query-only encoder's training images, so the evaluation tests generalization across concrete states rather than only across fresh render noise. Training and test pairs are balanced separately within each named action, and the same rendered panels are used for the trained and untrained encoder conditions. For standardized conditions, each feature coordinate is centered and scaled using probe-training data only.

The control in Table S3 changes the interpretation of the result. A randomly initialized encoder of the same architecture, never trained on any query or bit objective, is already highly predictive after standardization. Thus the probe detects action-relevant information that the continuous architecture preserves from the raw sensors; it does not identify query supervision as the source of that information. The query-trained encoder's held-out far-state depth-two query tensor is decoded perfectly in these runs, so the comparison is not driven by query inadequacy. Within the depth-two far class, the binary probe target is exactly the distinction revealed by the depth-three quotient: states with

**Table S1.** Per-seed optimized learned-soft conjunction accuracy with 95% Wilson binomial intervals over 12,000 evaluation rollouts. The exact-$V_2$ Bayes values are 91.80, 81.54, and 70.01%.

| Seed | $k = 1$ | $k = 2$ | $k = 3$ |
|---|---|---|---|
| 20260924 | 90.65 [90.12,91.16] | 84.31 [83.65,84.95] | 78.51 [77.76,79.23] |
| 20260925 | 89.87 [89.31,90.39] | 79.99 [79.27,80.70] | 69.91 [69.08,70.72] |
| 20260926 | 89.57 [89.02,90.11] | 81.03 [80.32,81.72] | 72.53 [71.73,73.32] |
| 20260927 | 89.82 [89.27,90.35] | 80.40 [79.68,81.10] | 71.34 [70.53,72.14] |
| 20260928 | 89.97 [89.42,90.49] | 83.52 [82.84,84.17] | 75.49 [74.71,76.25] |

**Table S2.** Query-only eight-step ablation. Five-seed mean class accuracy; the accompanying JSON contains seed-level values and $t_4$ intervals.

| Interface | $k = 1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Query-only soft | 86.95 | 77.69 | 68.00 | 59.10 | 51.49 | 46.50 | 44.38 | 44.60 |
| Query-only binary | 79.50 | 69.23 | 57.87 | 46.11 | 35.22 | 27.76 | 24.59 | 24.70 |

**Table S3.** State-disjoint decision probe. Intervals are five-seed $t_4$ intervals in the archived PyTorch 2.10 reference environment. "Std." means feature standardization fitted on the probe-training split only. The exact quotient rows are analytic under the balanced protocol.

| Condition | balanced accuracy (%) | AUROC |
|---|---|---|
| Query-trained sigmoid | 81.69 [79.84,83.55] | 0.891 |
| Query-trained sigmoid, std. | 82.11 [77.41,86.82] | 0.906 |
| Query-trained pre-sigmoid logits, std. | 86.94 [84.69,89.19] | 0.938 |
| Untrained random encoder, std. | 78.96 [72.26,85.67] | 0.925 |
| Exact depth-two quotient + action | 50.00 (analytic) | 0.500 |
| Exact depth-three quotient + action | 100.00 (analytic) | 1.000 |
| Raw eight-sensor render | 94.11 [92.86,95.36] | 0.987 |

exactly three missing controls are separated from the deeper far states, and the named action identifies whether the activated control is one of the three. Consequently the depth-two and depth-three quotient rows are analytic calibrations, not fitted measurements.

*Matched neural grounder with bit-head GRU and automaton controls.* All controls use the same jointly trained panel encoder, held-out glyph pools, and evaluation-rollout seeds as the Digits query-head GRU. They separate two comparisons that were previously conflated. First, the bit-head GRU keeps the same 64-unit GRU/action updater but initializes it from the encoder's eight sigmoid bit probabilities rather than the 37-dimensional depth-two query vector; this isolates the effect of the recurrent state information. Second, the exact and probabilistic automata receive that same bit-head information, so their comparison with the bit-head GRU isolates the transition-state implementation more cleanly. Table S4 gives the complete eight-step conjunction curve.

In the archived PyTorch 2.10 reference environment, retraining the matched controls reproduces the encoder metrics used by the corresponding scripts. Neural optimization and evaluation can shift modestly across CPU/PyTorch builds, so these seed intervals should not be read as cross-platform uncertainty bands.

**Table S4.** Five-seed mean conjunction accuracy on the Digits substrate. All rows use the same jointly trained panel encoder and identical evaluation-rollout seeds. Intervals are $t_4$ intervals in the archived reference environment.

| Interface | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ | $k=7$ | $k=8$ |
|---|---|---|---|---|---|---|---|---|
| Depth-two query-head GRU | 86.33 | 77.75 | 68.08 | 58.27 | 48.73 | 42.33 | 39.50 | 39.44 |
| Bit-head GRU | 92.51 | 90.86 | 89.49 | 88.80 | 88.75 | 89.28 | 90.03 | 90.81 |
| Exact automaton | 92.31 | 90.63 | 89.29 | 88.59 | 88.54 | 89.11 | 90.01 | 90.90 |
| Probabilistic automaton | 92.45 | 90.72 | 89.34 | 88.68 | 88.60 | 89.12 | 90.01 | 90.90 |

**Table S5.** Per-seed thresholded learned-partition diagnostics for conjunction. "Other reductions" counts exact-class reductions not explained by singleton classes merged into the far code.

| Seed | learned classes | total reductions | into far | other reductions | violations |
|---|---|---|---|---|---|
| 20260924 | 32 | 6 | 6 | 0 | 8 |
| 20260925 | 16 | 22 | 21 | 1 | 9 |
| 20260926 | 36 | 2 | 2 | 0 | 8 |
| 20260927 | 30 | 8 | 8 | 0 | 8 |
| 20260928 | 38 | 0 | 0 | 0 | 8 |

The bit-head GRU result is the critical ablation: its near-equality to the exact and probabilistic automata shows that, on this task, the large long-horizon gain over the query-head GRU is mainly due to retaining richer state information rather than replacing a GRU by symbolic transition code. Seed-level values and intervals are in `results/digits_bithead_gru_control.json` and `results/digits_same_grounder_automaton.json`.

The state-disjoint probe is reported only from the archived PyTorch 2.10 reference environment; its seed intervals quantify seed variation within that environment. Full per-seed values, split assignments, analytic baselines, and the random-encoder control are in `results/query_only_decision_probe.json`.

## S2. Intervention-resource supporting result

Consider the hypothesis class containing $F_0 \equiv 0$ together with $F_E(X) = \prod_{i \in E} X_i$ for unknown $|E| = r$. If at most $a$ targets can be changed per intervention and an episode contains at most $h$ interventions, then $ah < r$ makes all transcripts identical. If an episode can activate at most $r$ controls, exact worst-case identification requires $\binom{m}{r}$ episodes by direct hidden-hyperedge testing. With arbitrary pooled interventions, first activate all controls to distinguish $F_0$; if the output is one, complement tests satisfy

$$F_E(1_{[m]\setminus T}) = 0 \quad \Longleftrightarrow \quad T \cap E \neq \varnothing. \tag{S1}$$

A balanced recursive splitting procedure uses at most

$$1 + 2r\lceil \log_2 m \rceil \tag{S2}$$

experiments including the initial full-set check. Standard adaptive group-testing constructions can improve the dependence to $r \log_2(m/r) + O(r)$ in sparse regimes

**Table S6.** Degree-two extrapolation under bounded measurement noise ($10^5$ random perturbations per target size).

| $t$ | $\Lambda_{t,2}$ | adversarial $\lvert\mathrm{err}\rvert$ | random mean $\lvert\mathrm{err}\rvert$ | random 95th pct. |
|---|---|---|---|---|
| 3 | 7 | 0.007 | 0.00123 | 0.00298 |
| 5 | 31 | 0.031 | 0.00445 | 0.01069 |
| 10 | 161 | 0.161 | 0.02118 | 0.04822 |
| 20 | 721 | 0.721 | 0.09223 | 0.19757 |

**Table S7.** Scaling of the conjunction depth-two quotient and two sampling regimes.

| $m$ | $N_2(m)$ | $2^m/N_2(m)$ | uniform-far impurity | boundary impurity $3/m$ |
|---|---|---|---|---|
| 8 | 38 | 6.7 | 0.0959 | 0.375 |
| 12 | 80 | 51.2 | 0.01369 | 0.250 |
| 16 | 138 | 474.9 | 0.001606 | 0.1875 |
| 20 | 212 | 4946.1 | 0.000163 | 0.150 |

(Aldridge et al., 2019; see the main-paper references); exact monotone-DNF learning from membership queries provides a broader learning-theoretic context (Abasi et al., 2014; see the main-paper references).

## S3. Degree-two conditioning experiment

*Numerical robustness check.* To make the magnitude of the minimax factor concrete, we perturb every measured response by at most $\varepsilon = 10^{-3}$ for $r = 2$. The adversarial sign pattern from the proof attains $\varepsilon\Lambda_{t,2}$ exactly; as a non-adversarial comparison, Table S6 also reports $10^5$ independent draws with coordinate errors uniform on $[-\varepsilon, \varepsilon]$. This is a numerical illustration of conditioning, not evidence for the theorem. A fresh panel after 5, 10, 15, and 17 of the 20 target actions reduces the exact worst-case radius from $0.721$ to $0.391$, $0.161$, $0.031$, and $0.007$, respectively, illustrating the fresh-panel consequence of the theorem stated in Supplementary Section S6.3.

## S4. Scaling and closure-certificate details

The uniform-far column is structure agnostic. In the archived $m = 8$ sampled diagnostic, $n = 512$ uniform draws are taken independently within every non-singleton fiber/action pair. For the exact depth-two conjunction quotient this detects all eight violations with maximum empirical impurity $0.1211$. Repeating the same procedure on the thresholded learned conjunction partition for seed 20260924 also detects all eight of its exact violating pairs, with maximum empirical impurity $0.1055$. The boundary sampler is an oracle power demonstration that deliberately concentrates on missing sets of size three; it is not proposed as a generic learned-interface sampler. At $m = 20$, $n = 256$ detects all 20 action violations with empirical impurity range $0.1055$–$0.1758$ and mean $0.1520$. All three sampled outputs are generated by `code/certificate_sampling.py` and stored in `results/certificate_sampling.json`.

**Table S8.** Implementation check for the finite-core lifted updater. One hundred random rank-$\leq 2$ clutters on a fixed seven-control support are embedded in larger ambient domains. The canonical successor-orbit spectrum is identical across embeddings; the lifted compiler evaluates only the seven support actions plus one outside-support action type.

| Ambient $m$ | named action labels | lifted candidates | mean successor orbits |
|---|---|---|---|
| 8 | 8 | 8 | 4.52 |
| 16 | 16 | 8 | 4.52 |
| 32 | 32 | 8 | 4.52 |
| 64 | 64 | 8 | 4.52 |
| 128 | 128 | 8 | 4.52 |
| 256 | 256 | 8 | 4.52 |

**Table S9.** Mechanism recovery under exact and perceived panels. "Closure violations" is measured exhaustively on the learned residual partition; zero is predicted by the bounded-rank closure theorem and is therefore a sanity check.

| Channel | readings | exact recovery | depth-2 successor error | closure violations |
|---|---|---|---|---|
| Exact rank-$\leq 2$ panel | 1 | 100.00% | 0 | 0 |
| flip $p = .05$ | 1 | 24.68% | 12.27% | 0 |
| flip $p = .05$ | 3 | 83.16% | 2.63% | 0 |
| Digits parity | 1 | 51.36% | 7.39% | 0 |
| Digits parity | 3 | 94.72% | 0.76% | 0 |
| Digits parity | 5 | 99.68% | 0.05% | 0 |

The machine-readable run also records a maximum of eight successor orbits in every ambient domain. This is an implementation/scaling check of the finite-core theorem, not independent evidence for it.

## S5. Empirical mechanism and learned-partition details

The noiseless five-seed mechanism study contains 500 mechanisms per seed. It recovers all 2,500 rank-$\leq 2$ clutters exactly and makes zero errors in 250,000 subsequent transition checks. Enlarging the ambient class to rank three leaves a hidden triple for $99.4\%$ of mechanisms, with $19.88$ hidden alternatives per mechanism on average (standard deviation across seed means $0.54$); each sampled alternative produces the predicted cross-mechanism successor collision. The noisy-panel table is generated by `code/noisy_mechanism_learning.py` using the same five seeds. Each Digits read samples a fresh held-out glyph from the fixed held-out pool; errors are therefore image/classifier dependent rather than idealized i.i.d. flips.

The conjunction far class is never split by the thresholded code in these five runs; 0–21 exact singleton classes are instead merged into its dominant learned code. For rank two, the five learned class counts are 28, 28, 28, 29, and 30; three seeds match the exact 28-class quotient, while two split at least one exact class. These outputs, the exact-$V_2$ and exact-$V_3$ controls, training curves, and per-seed diagnostics are all stored in `results/lever_door_full_rerun.json`.

**Table S10.** Five-seed learned-partition diagnostics regenerated by the comprehensive lever-door rerun. Identity probes use the frozen soft query vector; chance levels are $1/219 = 0.46\%$ and $1/202 = 0.50\%$.

| Mechanism | identity probe | learned classes | exact violating pairs | learned-partition Bayes ($k = 1, 2, 3$) |
|---|---|---|---|---|
| Conjunction | 19.95% [14.01,25.89] | 16–38 | 8–9 | 88.84, 78.59, 67.06 |
| Rank two | 9.78% [5.87,13.69] | 28–30 | 0–3 | 100, 100, 100 |

## S6.1. Full proof: finite residual-core orbit completeness

**Proof.** (iii)⇒(i): By Lemma 13(a) of the main article, $R_j$ for $j \le h$ is determined by the truncated clutter. A permutation carrying $\mathcal{K}$ to $\mathcal{K}'$ therefore satisfies Equation (14) of the main article, and Corollary 5 applies. (i)⇒(ii) is trivial.

(ii)⇒(iii): Enumerate $U = \{u_1, \dots, u_n\}$ and let

$$\chi_{\mathcal{K}} := \exists x_1 \cdots \exists x_n \Big( \mathrm{Dist}(\bar{x}) \wedge \forall z \big( \mathrm{Supp}(z) \leftrightarrow \textstyle\bigvee_i z = x_i \big) \\ \wedge \bigwedge_{\substack{T \subseteq [n] \\ 1 \le |T| \le h}} \pm \mathrm{Min}_{|T|}(x_T) \Big) \wedge \pm R_0.$$

with signs recording whether $\{u_i : i \in T\} \in \mathcal{K}$ and whether $\emptyset \in \mathcal{K}$ (equivalently $F(S) = 1$). Its rank is at most $n + 1 + (h - 1) = n + h$, and it holds in $\mathfrak{M}_h(F, S)$ with $x_i = u_i$. By (ii) and Corollary 5 it holds in $\mathfrak{M}_h(G, S')$ with some witnesses $e_1, \dots, e_n$. There $\mathrm{Supp}$ defines the support of $\mathcal{K}'$, which therefore equals $\{e_1, \dots, e_n\}$. Every nonempty member of $\mathcal{K}'$ is a subset of this set of size at most $h$, and membership of each such subset, and of $\emptyset$, matches $\mathcal{K}$ by Lemma 13(b) of the main article. Thus $u_i \mapsto e_i$ carries $\mathcal{K}$ onto $\mathcal{K}'$. It extends to a permutation of $D$ because $|D \setminus U| = |D \setminus U'|$: both are $|D| - n$ when $D$ is finite, and both are $\aleph_0$ otherwise.

## S6.2. Full proof: finitary failure of orbit completeness

**Proof.** Let $G_1$ be a two-way infinite path on vertex set $D$ and $G_2$ a disjoint union of two two-way infinite paths on $D$. Let $F_i$ have clutter $\mathcal{H}_i = E(G_i)$, the edge set viewed as a family of 2-sets. In $\mathfrak{M}_{h'}(F_i, \emptyset)$, $R_0$ and $R_1$ are empty, and for $j \ge 2$, $R_j(\bar{a}) \leftrightarrow \bigvee_{\ell < \ell'} R_2(a_\ell, a_{\ell'})$, where $R_2$ is the edge relation of $G_i$. Each response structure is thus a quantifier-free definitional expansion of $G_i$, so it suffices that $G_1 \equiv G_2$.

Let $T$ be the theory of irreflexive symmetric graphs in which every vertex has exactly two neighbours and there is no cycle of any length $n \ge 3$. $T$ has no finite models, since a finite 2-regular graph contains a cycle. $T$ is $\kappa$-categorical for every uncountable $\kappa$: each connected component of a model is a two-way infinite path, hence countable, so a model of size $\kappa$ is a disjoint union of $\kappa$ such paths. By the Łoś–Vaught test (Hodges, 1993; see the main-paper references), $T$ is complete, and $G_1, G_2 \models T$. Finally, a permutation satisfying Equation (14) of the main article at depth 2 would be a graph isomorphism $G_1 \cong G_2$, which is impossible because only $G_1$ is connected.

## S6.3. Real-valued degree closure and minimax extrapolation

*Statement.* Let $F : 2^{[m]} \to \mathbb{R}$ have multilinear degree at most $r$, let $R_S = [m] \setminus S$, and define

$$c_U(S) = \sum_{V \subseteq U} (-1)^{|U|-|V|} F(S \cup V), \qquad |U| \le r.$$

Then: (a) $F(S \cup T) = \sum_{U \subseteq T,\ |U| \le r} c_U(S)$ and, after activating $i \in R_S$, $c_U(S \cup \{i\}) = c_U(S) + c_{U \cup \{i\}}(S)$, with coefficients above degree $r$ interpreted as zero; (b) the live state has $B_r(|R_S|) = \sum_{j=0}^{r} \binom{|R_S|}{j}$ coefficients, direct Möbius conversion costs $O(2^r B_r(|R_S|))$ arithmetic operations, and a fresh singleton update costs $O(B_r(|R_S|))$; (c) if $T \subseteq R_S$, $|T| = t > r$, and every order-$\le r$ measurement has additive error at most $\varepsilon$, the exact minimax error for estimating $F(S \cup T)$ over degree-$r$ functions is

$$\varepsilon \Lambda_{t,r}, \qquad \Lambda_{t,r} = \sum_{j=0}^{r} \binom{t}{j} \binom{t-j-1}{r-j};$$

(d) if $m' = |R_S| > r$ and $\varepsilon \Lambda_{m',r} \le \frac{1}{2}$, the same minimax equality holds with competitors constrained to $[0,1]$ on the whole cube; and (e) without the degree premise the deterministic omitted-tail error at target $T$ is exactly $\sum_{U \subseteq T,\ |U| > r} c_U(S)$. If $A \subset T$ is executed before collecting a fresh depth-$r$ panel at $S \cup A$, an estimator using only that fresh panel has exact minimax radius $\varepsilon$ when $s = |T \setminus A| \le r$ and $\varepsilon \Lambda_{s,r}$ otherwise.

*Proof.* Parts (a) and (b) are Möbius inversion on the Boolean lattice over the live coordinates $R_S$ and substitution of $t_i = 1$. For (c), substituting the Möbius coefficient formula into the reconstruction formula gives, for a measured $V \subseteq T$ with $|V| = j \le r$,

$$w_V = \sum_{\ell=0}^{r-j} (-1)^{\ell} \binom{t-j}{\ell} = (-1)^{r-j} \binom{t-j-1}{r-j}. \tag{S3}$$

Because $t \ge r+1$, the binomial coefficient in (S3) is positive for every $j \le r$, so $\operatorname{sign}(w_V) \in \{-1, +1\}$ is well defined. Thus the plug-in estimator has error at most $\varepsilon \sum_V |w_V| = \varepsilon \Lambda_{t,r}$.

The measurement coordinates $V \subseteq R_S$, $|V| \le r$, and the live coefficients $U \subseteq R_S$, $|U| \le r$, are equinumerous, and their zeta/Möbius map is square and invertible. Set $\xi_V = \varepsilon \operatorname{sign}(w_V)$ for measured $V \subseteq T$ and $\xi_V = 0$ for the other live measured coordinates. Let $D$ be the unique degree-$r$ function on the up-set of $S$ whose measured values are $\xi_V$. The worlds $D$ with noise $-\xi$ and $-D$ with noise $+\xi$ give identical observations but target values separated by $2\varepsilon\Lambda_{t,r}$, so every estimator incurs error at least $\varepsilon\Lambda_{t,r}$ on one world.

For (d), note first that $\Lambda_{t,r} \ge 1$, so the assumption $\varepsilon \Lambda_{m',r} \le \frac{1}{2}$ implies $\varepsilon \le \frac{1}{2}$. Also $\Lambda_{t,r}$ is nondecreasing for $t > r$: every summand $\binom{t}{j}\binom{t-j-1}{r-j}$ is nondecreasing in $t$. The same $D$ therefore satisfies $|D(S \cup A)| \le \varepsilon \Lambda_{|A|,r} \le \varepsilon \Lambda_{m',r}$ (with the direct bound $\varepsilon$ when $|A| \le r$). Hence $F_{\pm} = \frac{1}{2} \pm D$ lie in $[0,1]$ on the up-set. Extend them to the whole cube by ignoring already-active coordinates: $F^*_{\pm}(A) = F_{\pm}(S \cup (A \cap R_S))$. This preserves degree at most $r$ and the range bound. The same indistinguishable pair proves

the lower bound inside the globally $[0,1]$ class. Part (e) is the full Möbius expansion with degrees greater than $r$ separated.

## S6.4. Wasserstein–Lipschitz compositional bound

Let $(\Omega, \Sigma)$ be a measurable concrete state space, $(Z, d_Z)$ a metric abstract state space, and $e : \Omega \to Z$ a measurable encoder. Let $p_\alpha(w, b)$ and $\widehat{p}_\alpha(z, b)$ be concrete and abstract atomic scores, and let $K_{a,b}$ and $\widehat{K}_{a,b}$ be concrete and abstract action kernels. Interpret negation by $1 - x$, finite Boolean operations and quantifiers by min/max, and an action modality by expectation under the corresponding kernel. Assume uniformly that

$$|p_\alpha(w, b) - \widehat{p}_\alpha(e(w), b)| \le \varepsilon_0,$$

$$W_1\Big(e_{\#}K_{a,b}(w, \cdot), \widehat{K}_{a,b}(e(w), \cdot)\Big) \le \varepsilon_{\mathrm{tr}},$$
$$W_1\Big(\widehat{K}_{a,b}(z, \cdot), \widehat{K}_{a,b}(z', \cdot)\Big) \le \kappa d_Z(z, z').$$

and that every abstract atomic score is $L_0$-Lipschitz. With $\bar{\kappa} = \max\{1, \kappa\}$, every formula of intervention depth at most $h$ satisfies

$$\sup_{w,b} \Big| [\![\varphi]\!](w, b) - \widehat{[\![\varphi]\!]}(e(w), b) \Big| \le \min \left\{ 1, \varepsilon_0 + L_0 \varepsilon_{\mathrm{tr}} \sum_{j=0}^{h-1} \bar{\kappa}^j \right\}.$$

If $\kappa \le 1$, the bound is $\min\{1, \varepsilon_0 + hL_0\varepsilon_{\mathrm{tr}}\}$. For deterministic transitions, the $W_1$ transition premise reduces to the metric successor error.

**Proof.** Negation is an isometry and finite min/max are nonexpansive in the sup norm, so Boolean operations and quantifiers neither amplify semantic error nor increase the largest Lipschitz constant of their children. If an abstract subformula $\chi$ is $L$-Lipschitz, Kantorovich–Rubinstein duality and the kernel-contraction assumption above show that

$$z \longmapsto \int \widehat{[\![\chi]\!]}(z')\, \widehat{K}_{a,b}(z, dz') \tag{S4}$$

has Lipschitz constant at most $\kappa L$. A subformula with at most $j$ action layers therefore has Lipschitz constant at most $L_0\bar{\kappa}^j$. The maximum $\bar{\kappa} = \max\{1, \kappa\}$ is needed because min/max can combine branches of different action depths: a shallower branch may retain Lipschitz constant $L_0$ even when $\kappa < 1$.

At an action node, first replace the concrete child value at a successor by the abstract child value at its encoded successor, paying the inductive semantic error. Replacing the pushed-forward concrete successor law by the learned abstract law then costs at most $L\varepsilon_{\mathrm{tr}}$ in $W_1$. Summing along a branch of at most $h$ actions gives the displayed Wasserstein propagation bound; branching operations take maxima. For deterministic kernels, $W_1(\delta_x, \delta_y) = d_Z(x, y)$.

## S6.5. Full proof: finite-sample closure certificate

**Proof.** Singleton fibers cannot contain two states with different successors, so they contribute no non-congruence event. For a tested pair $(z, a)$ write $p_j = P_{z,a}(z_j)$. The probability that all $n$ sampled successors agree is $\sum_j p_j^n \le (\max_j p_j)^{n-1} \sum_j p_j = (1 - \eta_{z,a})^{n-1}$. If $\eta_{z,a} \ge \beta$, this is at most $(1-\beta)^{n-1}$. A union bound over the $K_*$ potentially nontrivial pairs proves Theorem 36 of the main article. For the exact statement, a non-congruent non-singleton fiber contains states with at least two distinct abstract successors. Atomic mass at least $\alpha$ places at least $\alpha$ probability outside the most likely successor class, so $\eta_{z,a} \ge \alpha$.

## S6.6. Boolean degree versus trigger rank

For a minimal activating set $E$ of a monotone Boolean mechanism, every proper subset of $E$ is false and $F(1_E) = 1$. Hence its Möbius coefficient is

$$\sum_{V \subseteq E} (-1)^{|E|-|V|} F(1_V) = 1,$$

so a degree-$|E|$ monomial is present. Thus real multilinear degree at most $r$ implies trigger rank at most $r$. The converse fails because $\mathrm{OR}_m$ has singleton minimal triggers but nonzero degree-$m$ coefficient.

## S6.7. Full proof: elementary equivalence of the star forests

Let $G_2$ be the disjoint union of countably many copies of $K_{1,n}$ for every $n \ge 1$, and let $G_1 = G_2 \sqcup K_{1,\omega}$. Both vertex sets are countably infinite, so both graphs can be placed on the pure domain $D$. We show $G_1 \equiv_q G_2$ for every $q$, where $\equiv_q$ is agreement on first-order sentences of quantifier rank at most $q$ in the language $\{E\}$.

*Stars.* $K_{1,\omega} \equiv_q K_{1,q}$. In the $q$-round Ehrenfeucht–Fraïssé game, Duplicator answers a centre by the centre, a previously chosen element by its partner, and a new leaf by a new leaf. At most $q$ leaves are ever needed, and $K_{1,q}$ has $q$ of them. Equality and adjacency are preserved, because leaves are adjacent only to the centre.

*Disjoint unions.* If $\mathfrak{A} \equiv_q \mathfrak{A}'$ and $\mathfrak{B} \equiv_q \mathfrak{B}'$, then $\mathfrak{A} \sqcup \mathfrak{B} \equiv_q \mathfrak{A}' \sqcup \mathfrak{B}'$. Duplicator runs her two winning strategies separately on the two summands and answers each move in the summand where it was made. Elements of different summands are distinct and non-adjacent on both sides, so the union of the two partial isomorphisms is a partial isomorphism.

*Conclusion.* Since $G_2$ contains infinitely many copies of $K_{1,q}$, removing one of them gives an isomorphic graph: $G_2 \cong K_{1,q} \sqcup G_2$. Hence

$$G_1 = K_{1,\omega} \sqcup G_2 \;\equiv_q\; K_{1,q} \sqcup G_2 \;\cong\; G_2.$$

As $q$ was arbitrary, $G_1 \equiv G_2$. For every depth $h'$, the response structures $\mathfrak{M}_{h'}(F_i, \varnothing)$ are quantifier-free definitional expansions of $G_i$: $R_0$ and $R_1$ are empty, and $R_j(\bar{x}) \leftrightarrow$

$\bigvee_{\ell<\ell'} E(x_\ell, x_{\ell'})$ for $j \geq 2$. A rank-$q$ sentence over $\sigma_{h'}$ therefore translates into a rank-$q$ sentence over $\{E\}$, and $(F_1, \varnothing) \equiv_{h'} (F_2, \varnothing)$ by Corollary 5 of the main article.

*Why the anonymous successor is unmatched.* In $\mathfrak{M}_2(F_1, \varnothing)$ the centre $c$ of $K_{1,\omega}$ realizes the partial type $\{\exists^{\geq n} y\, R_2(x, y) : n \in \mathbb{N}\}$, which every vertex of $G_2$ fails at some finite $n$. By Lemma 29 of the main article, this type fixes the depth-one class of the successor: $(F_1, \{c\})$ has infinitely many singleton residuals, while $(F_2, \{b\})$ has $\deg_{G_2}(b)$ of them. By Proposition 7 of the main article with $q > \deg_{G_2}(b)$, the two successors are $\equiv_1$-inequivalent.

## S6.8. Full proof: anonymous closure as nondeterministic abstraction

**Proof.** For Proposition 28 of the main article, (i)⇒(ii): the forth clause is the definition of the quotient relation. For the back clause, if $[\omega] \Rightarrow z'$ is witnessed by $\widetilde{\omega} \equiv_h \omega$ and $\widetilde{\omega} \to \widetilde{\omega}' \in z'$, anonymous closure yields a successor $\omega_a \equiv_h \widetilde{\omega}'$. Thus the quotient map is a bounded morphism. (ii)⇒(iii) is immediate. For (iii)⇒(i), let $\tau$ be an output-preserving bounded morphism with kernel $\equiv_h$. If $\omega \equiv_h \omega'$ and $\omega \to \omega_a$, forth gives $\tau\omega' = \tau\omega \Rightarrow \tau\omega_a$, and back gives $\omega' \to \omega'_b$ with $\tau\omega'_b = \tau\omega_a$, hence $\omega'_b \equiv_h \omega_a$. Forth and back force any admissible abstract relation to coincide with the quotient successor relation. The kernel of a bounded morphism is a bisimulation, and modal $\mu$-calculus formulas are bisimulation invariant.

## S6.9. Full proof: separation of anonymous closure, stabilization, and type-determinism

**Proof.** For part (a), let the trigger graph be $K_A \sqcup K_{B,C}$ with $|A| = r + 1$ and $|B| = |C| = r$. By Proposition 7 of the main article, a false state is classified by the number of singleton residuals, namely the number of neighbours of the activated set that would complete an edge. There are exactly three possibilities before acceptance. At the empty state the number is $0$, giving $Q_0$. A false state consisting of one clique vertex, or of any nonempty collection of vertices from only one bipartite side, has exactly $r$ singleton residuals and lies in $Q_r$. A false state containing one clique vertex together with one or more vertices from a single bipartite side has the $r$ remaining clique vertices together with all $r$ vertices on the opposite bipartite side as singleton residuals, giving $Q_{2r}$. Every other nonempty pattern completes an edge and lies in $T$. Direct inspection gives

$$Q_0 \Rightarrow \{Q_r\}, \qquad Q_r \Rightarrow \{Q_r, Q_{2r}, T\}, \qquad Q_{2r} \Rightarrow \{Q_{2r}, T\}, \qquad T \Rightarrow \{T\}.$$

Thus $\equiv_1$ is anonymously closed on this single-mechanism state space. Now activate $a \in A$ from the initial state. Its $r$ singleton residuals are the remaining vertices of $A$, while the hidden rank-two part is $K_{B,C} \cong K_{r,r}$. Activating $b \in B$ also gives exactly $r$ singleton residuals, namely $C$, but the hidden rank-two part is $K_A \cong K_{r+1}$. Hence the two successors are $\equiv_1$-equivalent and not $\equiv_2$-equivalent. For $r = 2$ these hidden cores are $C_4$ and $K_3$. Their ordinary successor sets agree, but their multiplicity profiles do

not: the clique-vertex and bipartite-side witnesses have sorted successor-count multisets $(1, 2, 4)$ and $(2, 2, 3)$, respectively (in the ordered classes $(Q_r, Q_{2r}, T)$ these are $(1, 4, 2)$ and $(2, 3, 2)$). Thus the separation is specific to the set-valued bisimulation induced by the ordinary existential modality, not to a graded counting modality.

For the smaller six-control witness, take

$$\mathcal{H} = \{\{0, 2\}, \{0, 3\}, \{0, 4\}, \{1, 3\}, \{3, 5\}, \{1, 2, 5\}, \{1, 4, 5\}\}.$$

An exhaustive response-structure calculation gives four depth-one classes and closure. The states $S = \{1\}$ and $S' = \{2\}$ both have one singleton residual. After that singleton is removed, the depth-two residual cores induce respectively a four-cycle and a triangle, so the states split at depth two. The companion checker verifies closure and the depth-two split directly. Its two witness multiplicity profiles are $(1, 1, 4)$ and $(1, 2, 3)$, so graded closure fails here as well. The exhaustive single-mechanism scan covers one through five controls and finds no closure-without-stabilization example at any positive horizon, proving six-control minimality for $h \geq 1$. At $h = 0$ minimality is different: the three-control clutter $\{\{2\}, \{0, 1\}\}$, corresponding to $X_3 \vee (X_1 \wedge X_2)$, is closed but not stabilized. The same scan finds no graded-closure-without-stabilization case through five controls, including all $1{,}050$ mechanism–horizon cases on five controls.

For (b), let $F$ be conjunction on a trigger $Q$ of size $h + 1$ and let the domain contain an additional control $d$. At the empty state every response relation of arity at most $h$ is empty, so a transposition of $q \in Q$ with $d \notin Q$ is an automorphism of the depth-$h$ response structure. But after activating $q$, the residual trigger has size $h$, while after activating $d$ it still has size $h + 1$; the successors are depth-$h$ inequivalent. On the other hand, every state has deficit at most $h + 1$, so the depth-$h$ and depth-$(h + 1)$ partitions coincide. Thus stabilization need not imply type-determinism.

For (c), Supplementary Section S6.7 proves elementary equivalence of the two star forests. A rank-two response structure at any depth is a quantifier-free definitional expansion of its depth-two structure: uniformly over the current state, $R_j(\bar{x}) \leftrightarrow R_0 \vee \bigvee_{\ell \leq \ell'} R_2(x_\ell, x_{\ell'})$ (with the repeated-index terms covering singleton residuals), so higher-arity response relations add no new first-order information beyond $R_0$ and $R_2$. Hence the depth-$h$ and depth-$(h + 1)$ equivalences agree for every $h \geq 2$ throughout the two-mechanism ambient space. Activating the centre of the added infinite star, however, produces infinitely many singleton residuals; every successor in the comparison mechanism has only finitely many. Proposition 7 of the main article separates those successors, so anonymous closure fails.

## S6.10. Full proof: FO(TC) compilation for the OR-DAG language

**Proof.** Only persistent positive interventions are used. Rename bound variables apart and carry the finite list $\pi = (t_1, \ldots, t_d)$ of intervention terms currently in scope. Under interventions on the elements denoted by $\pi$, the endogenous atom $X(y)$ is true exactly

when $y$ is equal to, or reachable from, at least one intervention root. Translate it to

$$\bigvee_{j=1}^{d} \big(y = t_j \vee [\mathrm{TC}_{u,v}\, E(u,v)](t_j, y)\big),$$

and translate baseline $X(y)$ with $d = 0$ to false. Equality and rigid edge atoms are unchanged; Boolean connectives and quantifiers translate homomorphically. An intervention modality appends its frozen term to $\pi$. If a later quantifier re-binds a variable occurring in $\pi$, alpha-rename that bound occurrence to a fresh variable before continuing, exactly as in the response-structure translation lemma. Induction on syntax eliminates every intervention modality and preserves the original free object variables.

## S6.11. Full proof: named query completeness

**Proof.** We prove the stronger uniform statement by structural induction: for every $d \le h$, every pair of states with equal $\mathsf{V}_d$, every common assignment $b$, and every open formula of intervention depth at most $d$, the two evaluations agree. This uniformity is needed because a do-modality moves the induction to successor states. We carry $b$ as part of the induction statement. A rigid atom $R(\bar{t})$ and an equality atom $t = t'$ agree because the domain, constants, rigid relations, and assignment are common. The atom $Y$ reads the coordinate $T = \varnothing$. Negation and conjunction preserve dependence on the same coordinates. For $\exists x\, \chi$, finite-domain semantics is a disjunction over $c_1, \dots, c_m$, and the induction hypothesis applies to $\chi$ under each extended assignment $b[x \mapsto c_i]$. For $[\mathrm{do}(X(t) = 1)]\chi$, the term $t$ resolves under $b$ to a named control $c_i$. If the whole formula has depth at most $h$, the induction hypothesis for $\chi$ needs only coordinates $F(1_{S\cup\{i\}\cup T})$ with $|T| \le h - 1$, all contained in the response tensor of Equation (7) in the main article.

Conversely, every coordinate $T = \{i_1, \dots, i_k\}$, $k \le h$, is read by

$$[\mathrm{do}(X(c_{i_1}) = 1)] \cdots [\mathrm{do}(X(c_{i_k}) = 1)]Y. \tag{S5}$$

A differing coordinate therefore distinguishes the states. Equation (9) in the main article is the set identity $S \cup \{i\} \cup T = S \cup (T \cup \{i\})$.

## S6.12. Full proof: rank-preserving response-structure translation

**Proof.** (a) Define $\psi^\pi$ by recursion on $\psi$:

$$(x = y)^\pi := (x = y), \qquad Y^\pi := R_\ell(y_1, \dots, y_\ell),$$
$$(\neg\psi)^\pi := \neg\psi^\pi, \qquad (\psi \wedge \chi)^\pi := \psi^\pi \wedge \chi^\pi,$$
$$([\mathrm{do}(X(x) = 1)]\psi)^\pi := \psi^{\pi x}, \qquad (\exists x\, \psi)^\pi := \begin{cases} \exists x\, \psi^\pi, & x \notin \pi, \\ \exists z\, (\psi[z/x])^\pi, & x \in \pi, \end{cases}$$

where $\pi x$ appends $x$ to $\pi$ and, in the last clause, $z$ is the first variable in a fixed infinite enumeration that occurs in neither $\psi$ nor $\pi$. A distinct fresh choice is made at each nested

rebinding, so the substitution is capture-free and $\psi[z/x]$ has the same depth, rank and size as $\psi$. The index of $R_\ell$ never exceeds $h$ because $\ell + \mathrm{dd}(\psi) \le h$ is preserved by every clause. No clause adds or removes a quantifier, so rank is preserved.

We prove Equation (15) of the main article by induction on the size of $\psi$, simultaneously for all $\pi$, $(F, S)$ and $b$. For $x = y$ both sides agree. For $Y$, the left side says $F(S \cup b[\pi]) = 1$, which is $R_\ell(b(y_1), \dots, b(y_\ell))$ by Definition 3 of the main article. Boolean connectives are immediate. For the do-clause, $(F, S \cup b[\pi]), b \models [\mathrm{do}(X(x) = 1)]\psi$ iff $(F, S \cup b[\pi x]), b \models \psi$, and the induction hypothesis for $\psi$ and $\pi x$ applies. For $\exists x\, \psi$ with $x \notin \pi$, changing $b(x)$ does not change $b[\pi]$, so the induction hypothesis applies to each $b[x \mapsto d]$. If $x \in \pi$, then $(F, S \cup b[\pi]), b[x \mapsto d] \models \psi$ iff $(F, S \cup b[\pi]), b[z \mapsto d] \models \psi[z/x]$ by renaming, because the state is a fixed set and $z$ does not occur in $\psi$. Since $z \notin \pi$, $b[z \mapsto d][\pi] = b[\pi]$, and the induction hypothesis for $\psi[z/x]$ applies. This last case is exactly where naive translation fails: a do-term must denote the element it denoted when the intervention was performed, even if its variable is later re-bound.

(b) Define $R_j(x_1, \dots, x_j)^\natural := [\mathrm{do}(X(x_1) = 1)] \cdots [\mathrm{do}(X(x_j) = 1)]Y$ (so $R_0^\natural = Y$), $(x = y)^\natural := (x = y)$, and let $\natural$ commute with connectives and quantifiers. Then $\mathrm{dd}(\theta^\natural) \le h$, rank is unchanged, and since each atom is evaluated from the current state $S$, a one-line induction gives the claim.

## S6.13. Full proof: finite anonymous orbit completeness

**Proof.** By Corollary 5 and Equation (14) of the main article, it suffices to show that two $m$-element $\sigma_h$-structures agreeing on rank-$m$ sentences are isomorphic. For an $m$-element structure $\mathfrak{A}$ with enumeration $d_1, \dots, d_m$, let $\delta_{\mathfrak{A}}$ be

$$\exists x_1 \cdots \exists x_m \Big( \bigwedge_{i<i'} x_i \neq x_{i'} \wedge \bigwedge_{\substack{T \subseteq [m] \\ |T| \le h}} \pm R_{|T|}(x_T) \Big),$$

where $x_T$ lists the variables indexed by $T$ in increasing order and the sign records whether $R_{|T|}$ holds of $(d_i)_{i \in T}$. Since each $R_j$ is symmetric and depends only on the underlying set, these literals determine the whole structure. A witness in an $m$-element structure enumerates its domain and therefore defines an isomorphism.

## S6.14. Full proof: exact depth-one rank characterization

**Proof.** $\mathfrak{M}_1(F, S)$ is the structure $(D, P_{F,S})$ with one unary predicate, expanded by the truth value $R_0 = F(S)$. For unary structures the $q$-round Ehrenfeucht–Fraïssé game is won by Duplicator iff each colour class has equal size in the two structures or size at least $q$ in both (Libkin, 2004, Ch. 3). Sufficiency: Duplicator answers an old element by its partner and a new element by a new element of the same colour, which exists because fewer than $q$ elements of that colour have been used. Necessity: if, say, $p := |P_{F,S}| < \min\{|P_{G,S'}|, q\}$, the rank-$(p+1)$ sentence $\exists x_1 \cdots x_{p+1}(\bigwedge_{i<i'} x_i \neq x_{i'} \wedge \bigwedge_i R_1(x_i))$ separates the structures, and likewise for complements. Corollary 5 of the main article transfers the result to the intervention language. For $q = \infty$, agreement at every finite $q$

forces equal cardinalities, and two unary structures on $D$ with equal class cardinalities are isomorphic.

## S6.15. Full proof: countable threshold lifting

**Proof.** In $\mathfrak{M}_h(F_k, S)$, $R_j(\bar{a})$ holds iff $|S \cup \{\bar{a}\}| \geq k$. If $d(S) > h$, then $|S \cup \{\bar{a}\}| \leq |S| + h < k$ for all $j \leq h$, so every $R_j$ (including $R_0$) is empty. If $d(S) = 0$, every $R_j$ is full. Hence all states with $d > h$ have literally the same response structure, and so do all states with $d = 0$. If $1 \leq d(S) = d(S') \leq h$, then $|S| = |S'|$, and any bijection $D \to D$ mapping $S$ onto $S'$ (which exists because both complements are countably infinite) satisfies Equation (14) of the main article. Thus equal values of $\min\{d, h+1\}$ give isomorphic response structures, and Corollary 5 of the main article gives equivalence at every rank.

Conversely, for $1 \leq j \leq h$ let $\vartheta_j := \exists x_1 \cdots \exists x_j \big( \bigwedge_{\ell<\ell'} x_\ell \neq x_{\ell'} \wedge R_j(x_1, \ldots, x_j) \big)$, and let $\vartheta_0 := R_0$. Since $D \setminus S$ is infinite, $\vartheta_j$ holds iff $d(S) \leq j$. These sentences have rank $j \leq h$. Their $\natural$-translations recover $\min\{d(S), h+1\}$.

For (ii), after $j$ applications of $\alpha$ the output is one exactly when $j \geq d(S)$, so words of length at most $h$ recover the same truncated deficit. For (iii), if $h \leq k-2$, the far class contains deficits $h+1$ and $h+2$ and one $\alpha$ maps them to the separated values $h$ and $h+1$; if $h \geq k-1$, all deficits $0, \ldots, k$ are distinct and $\alpha : d \mapsto \max\{d-1, 0\}$ is well defined.

For (iv), let $S \neq S'$ be finite and sub-threshold. If $|S| < |S'|$, activating $k - |S'|$ fresh elements outside $S \cup S'$ accepts from $S'$ but not from $S$. If $|S| = |S'| = s$, choose $u \in S \setminus S'$ and then activate $u$ followed by $k - s - 1$ fresh elements outside $S \cup S'$; the two resulting outputs differ. Thus infinitely many finite sub-threshold sets are pairwise Nerode-distinct when $k \geq 2$.

## S6.16. Full proof: residual characterization and truncation update

**Proof.** After an additional intervention $T$, the output is one exactly when $T$ contains some residual set $E \setminus S$. By Lemma 13 of the main article, among sets $T$ with $|T| \leq h$ the inclusion-minimal activating sets are exactly the truncated residual in Equation (12) of the main article. The truth table of all depth-$h$ intervention queries is determined by these minimal true sets and, conversely, the minimal true sets are recoverable from that truth table. The equivalence follows from Theorem 2 of the main article.

After action $i$, every residual $Q$ becomes $Q \setminus \{i\}$ and minimization removes dominated residuals. A residual dropped from the stored horizon because $|Q| \geq h+1$ yields $|Q \setminus \{i\}| \geq h$. Such a set cannot be a proper subset of any candidate retained at the new horizon, whose size is at most $h-1$. Hence omitted large residuals cannot change minimality among the candidates relevant to the $(h-1)$-truncation. Thus only the previously stored horizon-$h$ residuals are needed for Equation (25) of the main article.

## S6.17. Full proof: conjunction separation and recurrent lower bound

**Proof.** The class count follows from the residual description: every missing set of size at most $h$ gives a distinct response function, while all larger missing sets are depth-$h$ indistinguishable. For non-congruence when $h \leq m-2$, choose $M_1 \subset M_2$ with $|M_1| = h+1$ and $|M_2| = h+2$. They are in the same far class. If $i \in M_1$, action $i$ leaves missing-set sizes $h$ and $h+1$, which are distinguishable at depth $h$. Lemma 18 of the main article gives the failure of recurrent closure. At $h = m-1$, the only missing set larger than $h$ is $[m]$, so all $2^m$ concrete states are already separated and the quotient is Nerode.

For the recurrent lower bound, suppose two distinct reachable histories $S \neq T$ share an internal recurrent state after any earlier output feedback has been processed. Pick $i$ in their symmetric difference and relabel the pair if necessary so that $i \in S \setminus T$. If $S \neq [m]$, apply the controls in $[m] \setminus S$, ordering them arbitrarily except that one completion action is last. Both trajectories emit $Y = 0$ on all earlier continuation steps: the $S$ trajectory is not yet complete, while the $T$ trajectory always lacks $i$. Hence the feedback sequences remain identical and the internal states remain identical until the final prediction. On the last action the $S$ trajectory has $Y = 1$ and the $T$ trajectory has $Y = 0$, so their common prediction incurs error at least $1/2$ on one trajectory. If $S = [m]$, then $T$ is non-full; because $m \geq 2$, choose an action already active in $T$ (or, if $T = \varnothing$, any action that does not complete all controls). The full trajectory remains at output one while the non-full trajectory remains at zero, giving the same conclusion. Thus distinct concrete histories cannot share a recurrent state.

## S7. Verification details

Five verification scripts are supplied. `verify_theorems.py` checks all 168 four-control clutters (13,440 state–horizon cases), conjunction counts through $m = 8$, the interpolation grid, DAG/intervention cases, and hidden-trigger examples. `verify_structural.py` recomputes the certificate, threshold, and mechanism-learning identities. `verify_anonymous.py` runs 3,000 randomized response-structure formulas, the rebinding regression case, the small Ehrenfeucht–Fraïssé solver, and the exact-$V_2$ Bayes calculation. `verify_anonymous_closure.py` checks 6,107,808 hidden-layer update cases and the four-control pair scan (`-full` adds all non-isomorphic five-control pairs); `verify_single_mechanism_separation.py` checks Theorem 34(a), the six-control witness, and the positive-horizon no-example result through five controls. Checkers are read-only unless `-write` is supplied.

The archive also contains the five-seed neural, mechanism-learning, closure-sampling, and probe outputs. `results/reported_results.json` maps principal claims to scripts; scikit-learn Digits is loaded at runtime and is not redistributed.