# When Correct Predictions Require Incorrect States: A Compatibility Audit of Physical State Interfaces

**Anonymous authors**
Paper under double-blind review

## Abstract

Accurate predictions can require a state estimate outside its physical tolerance. We audit this conflict when several prediction queries share one state. A task-passing input shows that the queries can be satisfied. Ruling out task success throughout the allowed state neighborhood then certifies necessary distortion for every successful readout using that interface. On RoboMNIST, learned readouts pass the tasks on windows certified to require distortion, including frozen tests at other speeds and on another robot. The conflict persists with nonlinear heads that reduce MSE at the reference velocity by 28–35%. We then evaluate acceleration augmentation on 1,440 further windows. Relative to ordinary constant acceleration, mean full joint passing rises by 1.11–2.78 percentage points in three settings and falls by .28 points at medium speed, under the declared .5 m/s$^2$ acceleration tolerance. For released robot models, we certify which distance relations cannot be satisfied anywhere in the allowed joint neighborhoods. We then check the calibrated models at the same recorded inputs. At .01-degree joint and 1 mm distance tolerances, MUKCa's published calibrated models recover 98–100% of certified incompatible pairs on FR3 and two Panda robots, with weaker recovery on Kinova. The audit identifies conflicts that a readout change cannot resolve and tests whether changing the interface removes them.

## 1 Introduction

Prediction error scores a model at the supplied state. It leaves open whether any state within the physical tolerance can satisfy the prediction requirements. On 400 TACO windows, an affine head reduces recording-balanced MSE by 44.68% relative to constant velocity, yet the number admitting a state within both velocity and endpoint tolerances falls from 300 to 292. Thirty-two lose this joint solution and 24 gain one (Appendix G).

Consider a velocity estimate used by a constant-velocity head. During acceleration, shifting the estimate toward future motion can improve its endpoint prediction. The observed shift alone leaves two possibilities: training selected an inaccurate state even though an acceptable one exists, or every state that passes the task exceeds the velocity tolerance. A new readout may recover an acceptable state in the first case. The second requires changing the interface or its requirements.

We audit the interface by intersecting its task-feasible inputs with the allowed state neighborhood. If the intersection is empty but the tasks have a shared solution, we call the scene *distortion-required* (D). A joint witness establishes *feasibility* (F); tasks with no shared solution have *task inconsistency* (T). These categories depend on the interface, before a readout is chosen. An "incorrect" state in this paper exceeds the tolerance for a specified physical quantity and time definition. Figure 1 shows the classes and how they connect certification, learned outputs and recovery.

Task-based learning can favor intermediate outputs that help a downstream objective (Donti et al., 2017; Elmachtoub & Grigas, 2022). Our contribution is to certify when every task-feasible state must exceed tolerance, then check recovery on those same scenes after an interface change. On RoboMNIST, learned readouts realize the certified distortion under nonlinear heads; an acceleration extension restores joint success on further recordings. For released robot models, we bound the



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.

required joint change and identify distance relations that calibration recovers at unchanged recorded inputs. A laser-tracker case tests recovery against independent Cartesian measurements.

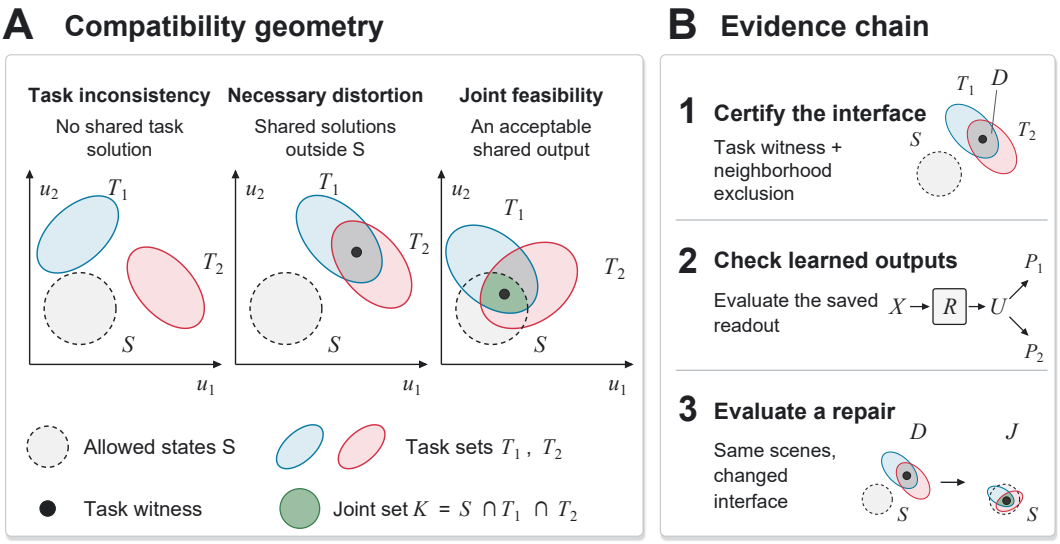


Figure 1: Compatibility geometry and the evidence chain (schematic). **A:** Task sets have no common output (T), intersect only outside the allowed state set $S$ (D), or admit a joint output in $K = S \cap T_1 \cap T_2$ (F). Dots are task witnesses; the dark green region is $K$. **B:** We certify the interface, evaluate learned outputs, and check actual joint recovery on the same scenes after an interface change.

## 2 Auditing a Physical State Interface

### 2.1 State and task constraints

Let $Z$ denote a scene, $s_Z$ its reference state and $X$ the information available to a readout $R$. The readout returns $U = R(X)$ in a nonempty state space $\mathcal{S}$. For query $a \in \mathcal{H}$, a fixed head $P_a$ predicts target $y_{Z,a}$. We define

$$S_Z = B_\epsilon(s_Z), \qquad T_{Z,a} = \{u : d_Y(P_a(u), y_{Z,a}) \le \tau_a\}, \qquad K_Z = S_Z \cap \bigcap_{a \in \mathcal{H}} T_{Z,a}. \tag{1}$$

State and task tolerances have physical units. Mixed-unit states use a separate tolerance for each coordinate group. For instance, the acceleration extension checks velocity and acceleration separately. An actual output passes jointly when it lies in $K_Z$. Compatibility asks whether $K_Z$ contains any output.

Each query can be feasible on its own while their intersection is empty: the interface must use one shared output for all queries.

Table 1: Interface classes. A task solution passes all queries with one shared state.

| Class | Any task solution? | Any joint solution? |
|---|---|---|
| F: jointly feasible | Yes | Yes |
| D: necessary distortion | Yes | No |
| T: task inconsistency | No | No |

### 2.2 Necessary distortion and its certificate

The minimum state change needed to pass the tasks is

$$\delta_{\text{req}} = \inf_{u \in \cap_a T_{Z,a}} d_S(u, s_Z), \qquad \inf \varnothing = +\infty. \tag{2}$$

A verified task witness shows that the tasks can be satisfied. A lower bound on $\delta_{\text{req}}$ above $\epsilon$ then certifies D for every successful readout using that head. The bound certifies a minimum required shift. The observed state error measures the shift made by a particular readout.

For a single constant-velocity query $P_h(v) = x_0 + hv$, $h > 0$, the task-feasible set is a ball centered at $(y - x_0)/h$ with radius $\tau/h$. Its distance from $v$ is

$$\Delta_\tau(v, y; P_h) = (\|x_0 + hv - y\|_2 - \tau)_+ / h. \tag{3}$$

For several queries, $\delta_{\text{req}}$ measures distance to their common intersection. The set can disappear when another query is added, as in Figure 1A (T).

The numerical audit checks candidate inputs for task or joint feasibility and uses bounds over the entire semantic set to establish exclusion. For affine heads, feasible points and dual separation margins resolve F, D and T in the reported trajectory experiments. Interval and affine-envelope bounds cover the semantic ball for learned ReLU heads. Directional Taylor bounds cover full angle boxes for the external robot configurations and their geometric relations. The appendices give each bound, its numerical guards and independent implementation checks.

Incomplete evidence receives label U. A failed witness search leaves task feasibility unresolved; failed exclusion alone leaves joint feasibility unresolved. Each head and scene has one label shared by all compared readouts. Candidate outputs can supply a witness. The exclusion bound still covers every input in the specified state neighborhood.

The classes also give a readout-independent failure bound. Since no output can pass jointly on D or T,

$$\Pr\{R(X) \notin K_Z\} \geq \Pr(D) + \Pr(T). \tag{4}$$

Task failure alone is at least $\Pr(T)$. On feasible scenes, ambiguity in the observations or an inaccurate readout can still cause failure. The bound isolates the error that persists for every readout under the fixed contract. Appendix B gives the proof and the geometric checks.

### 2.3 Checking an interface change

After changing the head, we check which scenes gain an acceptable state and whether the evaluated readout returns one. On an old D scene, a new joint witness certifies $D \to F$. A saved output satisfying the complete new contract establishes actual joint success, denoted $D \to J$:

$$n_{D \to J} = \sum_i \mathbf{1}\{L_i^{\text{old}} = D\} \mathbf{1}\{U_i^{\text{new}} \in K_i^{\text{new}}\}. \tag{5}$$

Here $L_i^{\text{old}}$ is the old certified class. We also record losses on previously F scenes, since an aggregate increase in feasibility can conceal such losses.

Comparisons preserve the required quantities and their time definitions. When the state is expanded from velocity to velocity and acceleration, the old velocity head is evaluated on that same expanded contract and ignores the acceleration coordinates. For the external robot interfaces, the evaluated states are the recorded joint configurations. A passing calibrated prediction gives a new F witness at unchanged input. The relation audit treats each record pair as one instance, with the two configurations concatenated into its state.

Paired reference checks use one shared error set. Old semantic exclusion, an old task witness and new joint success must all hold throughout it; losing a certificate leaves the result unresolved. Published reference trajectories supply offline scoring and certification targets, not future inputs to learned readouts.

# 3 Necessary Distortion in Learned Readouts

## 3.1 State, inputs and training objectives

On RoboMNIST (Behzad et al., 2025), we evaluate frozen affine readouts on 24 recordings from two activities excluded from head fitting and readout development (720 windows). The readouts use past feedback about the robot's end-effector origin. Its backward 0.1-second velocity is the state, and the shared head predicts displacement at 0.1 and 0.2 seconds. The semantic tolerance is 0.1 m/s and each endpoint tolerance is 20 mm. Published trajectories supply the scoring references. Inputs use only feedback available by the prediction time (Appendix C).

Each affine readout uses two past velocities to correct the current estimate. State supervision fits the correction to velocity; task training fits it to the two endpoints. Validation selects the regularization and correction scale. The joint-selected control chooses from a frozen readout library by joint passing. Appendix C.1 specifies these fits and the additional controls.

## 3.2 Learned success on certified D windows

The audit finds 527 F, 190 D and three T windows. A task-trained readout passes both queries on 18 of the 190 D windows, compared with zero for the state-supervised readout (Table 2). All 18 have a necessary-shift lower bound above 0.1 m/s, and their actual outputs exceed that tolerance. These successes occur in both activities and 16 of 24 recordings. Figure 3 pairs each required-shift lower bound with the actual shift and task margin.

The recording/activity-balanced D-task success rate is 9.73%, with conditional 95% interval $[6.71, 12.71]$ points above state supervision (Appendix C.1).

Table 2: Frozen RoboMNIST readouts on 720 held-out windows. Rates are percentages. All methods share past inputs and the physical head. D-task success counts successes among the 190 windows requiring distortion. Appendix C.1 gives all baselines, including set-target, and state RMSE.

| Readout | Semantic failure ↓ | Task pass ↑ | Joint pass ↑ | D-task success |
|---|---|---|---|---|
| Past velocity | .56 | 49.86 | 49.86 | 0/190 |
| State-supervised | .56 | 49.58 | 49.58 | 0/190 |
| Joint-selected | 3.19 | 56.11 | 55.42 | 1/190 |
| Task-trained | 25.28 | 64.17 | 52.50 | 18/190 |

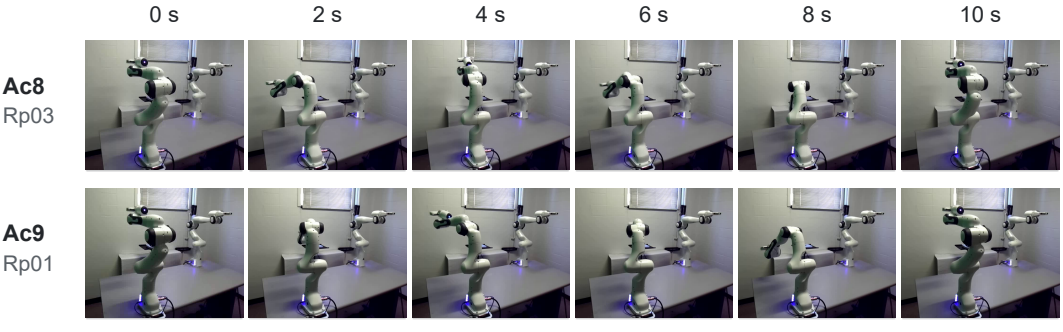


Figure 2: Robot motion in two held-out RoboMNIST recordings (robot 1/high speed): activity 8, repetition 3, and activity 9, repetition 1. Each Rx1 video's left lens is shown every two seconds from 0 to 10 seconds with one fixed crop. These frames illustrate the scene; recorded feedback supplies readout inputs and audit references. Reproduced from Behzad et al. (2025), CC BY 4.0.

Task passing rises by 14.58 percentage points relative to state supervision, while semantic failure rises from 0.56% to 25.28%. Joint passing remains below the joint-selected baseline; the joint gain over state supervision has a 95% interval spanning zero. Of the 105 net additional task successes, 18

occur on D and 87 on F (Appendix C.1). Task training thus incurs state error on feasible windows as well as on those requiring distortion.

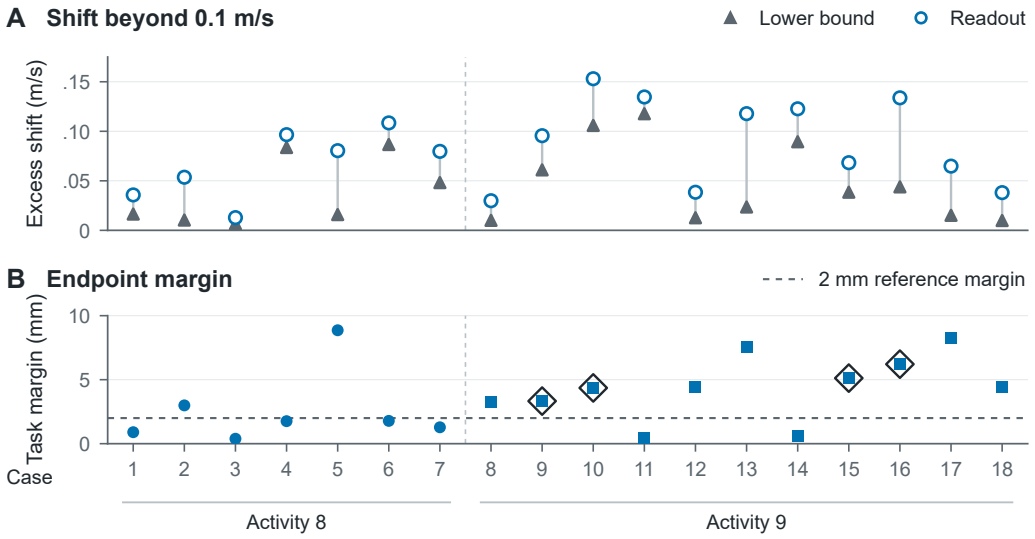


Figure 3: All 18 nominal D successes. A: state shift minus the 0.1 m/s tolerance; positive values exceed the state tolerance. Triangles show lower bounds on the required shift; circles show actual shifts. B: 20 mm minus maximum endpoint error; positive values pass both tasks. Diamonds mark four cases where D and the saved output's task success remain certified under assumed 1 mm position errors (dashed threshold: 2 mm). Cases are ordered by activity, recording and time.

For these 18 successes, both necessary distortion and actual task passing remain certified under the stated reference-error sets up to position radii of .193–2.178 mm (Appendix C.2).

### 3.3 More accurate heads and frozen transfer

We replace the affine head with a learned nonlinear map, while state-, task- and joint-trained readouts share a $6 \to 32 \to 3$ MLP and the same past inputs. We train on activities 0–3 of robot 1 at high speed, select on activities 4–5, and refit on activities 0–5 (Appendix C.3). The three readout objectives minimize normalized velocity error, mean endpoint error, and the mean of all three errors, respectively. Validation selects by state RMSE, task passing or joint passing to match the objective.

The nonlinear head has the form

$$P(u) = \begin{bmatrix} .1u \\ .2u \end{bmatrix} + .02\, g_\theta(u/.1),$$

where $g_\theta$ is a $3 \to 32 \to 6$ ReLU network correcting the CV displacements. Development-set MSE at the reference velocity selects width 32 from 16/32, regularization and training duration; the three fitted seeds share this selection.

We first evaluate the three fitted seeds on the previously inspected S3 windows. We then freeze all models and test robot 1 at medium and low speed and robot 2 at high speed, using activities 8–9, repetitions 1–12 in each setting. These three groups add 72 recordings and 2,160 windows. State definitions, input construction and tolerances remain fixed; the transfer tests use each robot's published coordinates without adaptation.

When given the reference velocity, the nonlinear heads reduce MSE by 28.26–34.67% relative to the affine baseline. Task-trained readouts succeed on certified D windows for all three head seeds in all four settings (Table 3).

On S3, the nonlinear heads certify 17–21 more feasible windows than the affine head. The successes span distinct physical windows and recordings; Appendix C.4 gives the independent checks and Appendix C.5 reports their distribution.

Table 3: Nonlinear RoboMNIST heads and readouts, 720 windows per setting. MSE reduction is relative to the affine head at the reference velocity; joint passing compares readout objectives. Both metrics average three seeds. Seed 1/2/3 denote 40101/40102/40103. Each fraction counts task-trained successes among that head's certified D windows. U is unresolved, not T. S3 is retrospective; the other settings test frozen models on newly accessed recordings.

| | Head quality | Joint passing (%) | | | Certified D windows | | | Unresolved |
|---|---|---|---|---|---|---|---|---|
| | MSE | Training objective | | | Task successes / certified D | | | U (%) |
| Setting | reduction (%) | State | Task | Joint | Seed 1 | Seed 2 | Seed 3 | range |
| R1 high (S3) | 34.23 | 56.90 | 54.86 | 66.53 | 12/16 | 7/10 | 2/9 | 21.94–23.06 |
| R1 medium | 31.32 | 60.83 | 57.64 | 68.47 | 3/3 | 3/3 | 1/1 | 17.36–18.19 |
| R1 low | 28.26 | 72.27 | 59.77 | 75.19 | 2/2 | 3/4 | 1/2 | 8.47–9.72 |
| R2 high | 34.67 | 57.45 | 57.45 | 68.24 | 15/16 | 10/12 | 5/8 | 19.44–20.14 |

Task training improves task passing over state training on S3, medium speed and robot 2/high speed (Figure 5 in Appendix C.5). At low speed the difference is $-.83$ points, with conditional 95% interval $[-2.45, .69]$. This setting therefore has certified D successes but no clear overall task-passing gain. Joint training has the highest mean joint passing in every setting (Table 3).

All readouts share one F/D/U label for each head and window. A verified task output establishes task feasibility; a separate global certificate excludes the entire semantic ball. Certification uses the exported heads and nominal published references (Appendix C.4).

## 4 Repairing the Velocity Interface

We extend the velocity interface and evaluate 48 further held-out recordings: activities 8–9, repetitions 25–30, in the same four robot/speed settings. Each setting has 360 windows, giving 1,440 in total. The five original candidate heads and their readouts are frozen before evaluation; Table 4 also includes the subsequently added data/budget-matched nonlinear control.

### 4.1 A head matched to backward-window velocity

The extended state is $z = (s_w, a_w)$, where $w = .1$ s and $a_w = [s_w(t) - s_w(t-w)]/w$. Under constant acceleration, the window-matched head is

$$P_h^{\mathrm{CA}}(z) = hs_w + \tfrac{1}{2}h(h+w)a_w.$$

The identity $s_w = v(t) - wa_w/2$ gives this coefficient and preserves the state's backward-window meaning. We also fit a $6 \to 32 \to 6$ residual with componentwise $.005\tanh(\cdot)$ output, bounding each endpoint-coordinate correction by 5 mm. Ordinary CA uses $h^2/2$ in place of $h(h+w)/2$. Every readout has the same $6 \to 32 \to 6$ architecture and receives the same two past velocity estimates. Full joint passing checks velocity within .1 m/s, acceleration within .5 m/s$^2$ and both endpoints within 20 mm. Velocity-only heads ignore the acceleration coordinates, while their readouts undergo all four checks. Appendix D gives training and selection details.

### 4.2 Joint passing and paired recovery

Relative to ordinary CA, one of the five original candidates, the residual extension raises mean full joint passing by 2.78, 1.11 and 2.50 points in R1 high, R1 low and R2 high, respectively. It lowers passing by .28 points in R1 medium. These means use the .5 m/s$^2$ acceleration tolerance fixed before confirmation evaluation. Joint validation on development records selects the residual head for all three seeds.

The retrospective nonlinear control uses the same training data and budget as the residual head. The residual head gains 1.85–5.65 points across the four settings; the low-speed interval crosses zero. Tightening only the acceleration tolerance to .25 m/s$^2$ reverses all four comparisons with this control (Appendix D.1).

Table 4: Full joint passing (%) on 360 new windows per setting, averaged over three seeds. Joint-trained $6 \to 32 \to 6$ readouts are scored on velocity, acceleration and both endpoints. Bold marks the largest row mean. The frozen nonlinear head is retained in Table 12; the control shown here matches the larger training set and budget of the residual head. †: added after inspecting the initial confirmation results, using the same 48 records.

| Setting | Original affine | Data/budget-matched nonlinear† | Ordinary CA | Window-matched CA | Window-matched CA + residual |
|---|---|---|---|---|---|
| R1 high | 52.22 | 65.83 | 68.70 | 69.17 | **71.48** |
| R1 medium | 60.28 | 70.00 | **73.43** | 73.06 | 73.15 |
| R1 low | 70.19 | 75.65 | 76.39 | 74.91 | **77.50** |
| R2 high | 55.93 | 67.69 | 70.09 | 70.56 | **72.59** |

Table 5: Restored feasibility and actual joint success on the same windows. $D \to F$ gives the range across seeds; $D \to J$ lists saved readout outputs passing the complete new contract for seeds 1/2/3. $F \to D$ is the same for all three seeds.

| Setting | Old D | $D \to F$ | Actual joint success $D \to J$ | $F \to D$ |
|---|---|---|---|---|
| R1 high | 96 | 78–84 | 60/52/60 | 8 |
| R1 medium | 52 | 44–45 | 29/28/29 | 3 |
| R1 low | 22 | 14 | 11/11/11 | 3 |
| R2 high | 93 | 78–81 | 52/51/52 | 3 |

**Recovery on certified D windows.** On R1 high, the residual heads change 78–84 of the old affine head's 96 D windows to F. Their saved joint-trained outputs pass all requirements on 60, 52 and 60 of those original D windows (Table 5). The $D \to J$ counts record actual passing outputs on windows where the old interface had no joint solution under the same contract. Some old F windows become D after the change. Appendix D reports distinct-window counts and all per-seed transitions, including unresolved cases (Table 11).

We check old D and new joint success under the same reference errors. Independent .1 mm balls on five reference positions preserve both on 50–56, 27–28, 9 and 49 windows per seed in the four settings. Appendix D.2 gives the complete 0–2 mm scan under this assumed error model.

## 5 Auditing Published Robot Interfaces

The learning experiments test whether trained readouts realize the certified distortion and recovery. The robot experiments hold recorded states fixed and compare original and calibrated models to identify conflicts removed by calibration.

### 5.1 Geometric relations at unchanged joint states

For MUKCa's released robot models (Franzese et al., 2026), we certify which distance relations the original model cannot satisfy anywhere in the allowed joint neighborhoods. We then evaluate the calibrated model at the same recorded inputs (Figure 4). The fixture has two sockets, each supporting the same tool-mounted ball at a fixed center. Configurations in one socket should predict the same center; configurations in different sockets should predict centers 50 mm apart.

A cross-socket scene is a pair $(q_i, q_j)$ with 12 or 14 joint coordinates, six or seven per configuration. The scalar head is $P(u, v) = \|f(u) - f(v)\|_2$, with target 50 mm and tolerance 1 mm. All in-chain joints in both configurations may vary simultaneously within .01 degrees of their recorded values:

$$\max\{\|u - q_i\|_\infty, \|v - q_j\|_\infty\} \le \epsilon_q, \qquad \epsilon_q = .01^\circ.$$

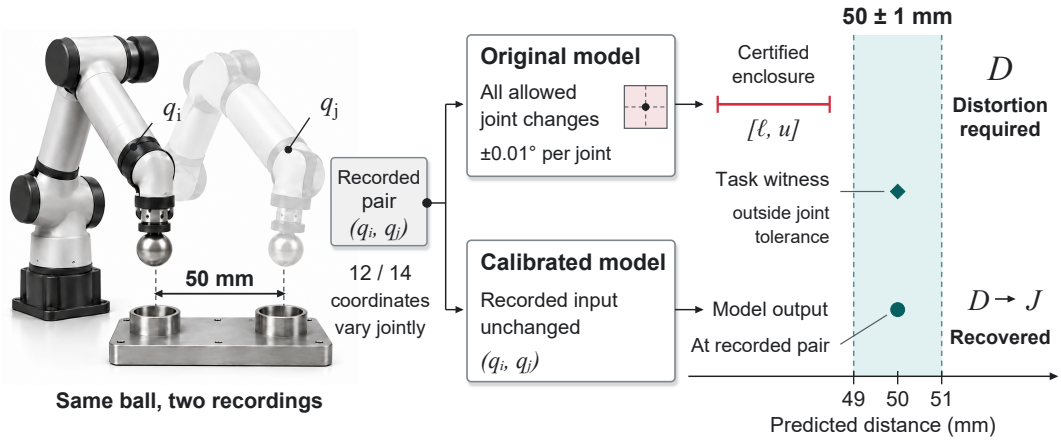


Figure 4: Cross-socket audit of a recorded configuration pair (schematic). The same ball occupies two sockets with 50 mm center separation. The old-model enclosure covers all allowed changes in both configurations; exclusion and a separate task witness certify D. The released calibrated model passes at the unchanged recorded pair (J). Both rows share the distance axis and the 49–51 mm task band. Geometry and interval positions are illustrative.

These tolerances are the declared audit requirements. A global Taylor enclosure covers the entire box. If it excludes the task interval and a separate task witness passes, the pair is D: any task-passing input pair must move at least one joint coordinate beyond tolerance. Expanding the box gives a lower bound on the necessary maximum-coordinate change $\delta_{\text{req}}$.

We include FR3, Panda 2, Panda 3, KUKA iiwa 14 and Kinova Gen3lite: all five systems with both released models, raw records and fitting KPIs. The main summary uses 14 locations and 838 records, excluding the five locations listed in those KPIs. This is a retrospective audit of released models using the published metadata split. The available metadata leave the records' use during other stages of original model development unknown.

Table 6: Cross-socket audit of original and published calibrated models. Passing uses unchanged records; recovered pairs were certified D under the old model. Lost counts old passes that now fail. Median bound is the certified lower bound on necessary joint change among old D pairs, in degrees.

| System | Pairs | Old→new passes | Recovered / certified D | Lost | Median bound (degrees) |
|---|---|---|---|---|---|
| FR3 | 2,444 | 194→2,398 | 2,080/2,117 | 2 | 0.0878 |
| Panda 2 | 5,093 | 888→5,093 | 3,735/3,735 | 0 | 0.0473 |
| Panda 3 | 2,352 | 159→2,352 | 2,107/2,107 | 0 | 0.1475 |
| KUKA | 1,745 | 1,432→1,745 | 69/69 | 0 | 0.0126 |
| Kinova | 1,282 | 211→335 | 211/1,019 | 107 | 0.1060 |

The calibrated models pass at unchanged recorded inputs for 98.25%, 100% and 100% of old D pairs on FR3 and the two Pandas. Their median necessary-radius lower bounds are .0878, .0473 and .1475 degrees. KUKA passes most pairs before calibration and has 69 certified D pairs. Kinova recovers 211 of 1,019 while losing 107 old passes. At a stricter .5 mm distance tolerance, FR3 and the Pandas recover 82.0%, 97.2% and 85.7% of their respective D pairs.

Each certificate concerns a scalar distance for one recorded pair. Absolute position and pose accuracy require separate checks. Pairwise witnesses establish feasibility for each pair without establishing a shared correction map, and pairs that share records are statistically dependent. Appendix E gives all 14 locations, same-socket results and reference sensitivity. MUKCa provides the calibration

procedure and models. Our audit identifies recovered relations that the original model could not satisfy anywhere in the allowed neighborhoods.

### 5.2 Independent Cartesian measurements

The CTU Motoman/Leica data (Stepanova et al., 2022) provide an external Cartesian reference for a seven-angle interface. We fit registration and calibration on 465 poses and reserve 121 by a fixed pose-ID rule within the same acquisition. At .01 degrees per joint and 1 mm position tolerance, the full-box certificate and task witnesses identify 99 old D poses. Calibration restores success at unchanged recorded inputs on 12 of them; 11 retain both old necessity and new success under reference balls of radius $3U95_i$, scaled from the reported measurement uncertainty.

Total recorded-input passing rises from 7 to 16, including six old passes that are lost. Reoptimizing registration alone leaves seven passes. The paired audit identifies 12 gains on poses with certified conflicts under the original interface. This case tests recovery against measured Cartesian positions, complementing the fixture-distance relations above. Appendix F gives the model, full transition table and conditional reference-error checks.

## 6 Related Work

Task-based learning and Smart Predict-then-Optimize train intermediate predictions for downstream use (Donti et al., 2017; Elmachtoub & Grigas, 2022). Value-equivalent models preserve selected Bellman updates (Grimm et al., 2020). These objectives explain why a task-trained state can depart from its supervised target. We examine when the declared tasks require every successful state to exceed its tolerance.

Concept bottlenecks and reasoning shortcuts study predictive success with altered intermediate meanings (Margeloiu et al., 2021; Marconato et al., 2023); identifiability results also address the inference layer (Bortolotti et al., 2025). Our state and task tolerances quantify that departure in physical units. Checking task-only feasibility also distinguishes necessary distortion from incompatible tasks that no shared state can satisfy.

Backward error measures the input change needed to explain an output (Higham, 2002). Model invalidation and perception contracts test consistency with error sets (Harirchi & Ozay, 2018; Astorga et al., 2023; Sun et al., 2024; Li et al., 2025). We apply feasible witnesses and separation bounds to a state shared across tasks, and evaluate saved outputs on the same scenes after changing the interface. Robot calibration estimates the map from recorded joints to Cartesian predictions (Stepanova et al., 2022; Franzese et al., 2026). We audit these authors' models by ruling out task success throughout the old model's allowed state neighborhoods, then checking recovery under the calibrated model at unchanged inputs.

## 7 Discussion and Conclusion

A readout's state error can reflect a conflict in the interface itself. Our audit certifies that conflict and bounds the state change needed to satisfy the tasks. RoboMNIST readouts realize the certified distortion under more accurate nonlinear heads, and the acceleration extension recovers joint success on D windows in further recordings. For published robot models, calibration recovers distance relations that the original model cannot satisfy anywhere within the allowed joint neighborhoods. The laser-tracker case checks recovery against an independent position reference.

The diagnosis depends on the declared physical quantities and tolerances. Reference-error sets specify separate measurement assumptions. Incomplete certificates remain unresolved. The paired tables show which gains survive these reference errors and retain old successes lost after a model change, allowing recovery to be assessed alongside its cost.

For model development, a D certificate rules out repairing the readout alone under the current contract. Evaluating a candidate interface on the same instance checks both joint feasibility and actual joint success. Paired recovery and loss counts show which conflicts the change removes and which old successes it loses, complementing average prediction and task scores.

### AI USE STATEMENT

Generative AI assisted study design, code, analysis, and writing; the authors take responsibility for the paper.

### REPRODUCIBILITY STATEMENT

An anonymous code ZIP is included in the supplementary material, with evaluation scripts, model parameters, compact evaluation inputs and data-download instructions. Its README provides commands for score reproduction and certificate checks. The supplement specifies state definitions, task tolerances, data partitions, fitted models and numerical checks. The supplementary results include readout outputs, task witnesses, exclusion certificates, all seeds and complete tolerance grids. RoboMNIST S3 is reused retrospectively for the nonlinear study; its three transfer groups use models and recording lists frozen before evaluation. The acceleration study fixes five candidates, the .5 $m/s^2$ acceleration tolerance and the .25/.5/1.0 $m/s^2$ sensitivity grid before evaluating 48 new recordings, as documented by the evaluation protocol and its pre-evaluation hash. Its data/budget-matched nonlinear control is added after inspecting those results, using the same recordings. CTU registration and calibration use only the development poses. Its Taylor bound refines the initial verifier without changing models or thresholds, and the registration-only control follows the pilot. TACO models precede test-trajectory access. The MUKCa audit fixes the public models, eligible systems, raw-row policy and tolerance grid before computing its scores; it uses existing records, with the locations listed in fitting KPIs separated from the main summary. Tolerance and reference analyses reuse these frozen results.

## REFERENCES

Angello Astorga, Chiao Hsieh, P. Madhusudan, and Sayan Mitra. Perception contracts for safety of ML-enabled systems. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA2), 2023. doi: 10.1145/3622875. Article 299, 28 pages.

Kian Behzad, Rojin Zandi, Elaheh Motamedi, Hojjat Salehinejad, and Milad Siami. RoboMNIST: A multimodal dataset for multi-robot activity recognition using WiFi sensing, video, and audio. *Scientific Data*, 12:326, 2025. doi: 10.1038/s41597-025-04636-2.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In *Advances in Neural Information Processing Systems*, volume 38, pp. 173204–173243, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fcdccd419c4dc471fa3b73ec97b53789-Abstract-Conference.html.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Franka Emika. Robot arm & control: Data sheet, 2020. URL https://download.franka.de/Datasheet-EN.pdf. Release version: April 2020.

Giovanni Franzese, Max Spahn, Jens Kober, Javier Alonso-Mora, and Cosimo Della Santina. Accurate and affordable cobot calibration without external measurement devices. *Communications Engineering*, 5:138, 2026. doi: 10.1038/s44172-026-00633-4.

Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Farshad Harirchi and Necmiye Ozay. Guaranteed model-based fault detection in cyber-physical systems: A model invalidation approach. *Automatica*, 93:476–488, 2018. doi: 10.1016/j.automatica.2018.03.040.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, second edition, 2002. doi: 10.1137/1.9780898718027.

Yangge Li, Chenxi Ji, Jai Anchalia, Yixuan Jia, Benjamin C. Yang, Daniel Zhuang, and Sayan Mitra. Lyapunov perception contracts for operating design domains. In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 1053–1065, 2025. URL https://proceedings.mlr.press/v283/li25c.html.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_TACO_Benchmarking_Generalizable_Bimanual_Tool-ACtion-Object_Understanding_CVPR_2024_paper.html.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72507–72539, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e560202b6e779a82478edb46c6f8f4dd-Abstract-Conference.html.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. URL https://arxiv.org/abs/2105.04289.

Karla Stepanova, Jakub Rozlivek, Frantisek Puciow, Pavel Krsek, Tomas Pajdla, and Matej Hoffmann. Automatic self-contained calibration of an industrial dual-arm robot with cameras using self-contact, planar constraints, and self-observation. *Robotics and Computer-Integrated Manufacturing*, 73:102250, 2022. doi: 10.1016/j.rcim.2021.102250.

Dawei Sun, Benjamin C. Yang, and Sayan Mitra. Learning-based inverse perception contracts and applications. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11612–11618, 2024. doi: 10.1109/ICRA57147.2024.10610329.

Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGIdiRqtm.

# A State Requirements and Reference Measurements

Table 7 distinguishes required prediction/state errors from uncertainty in the scoring references. All primary values retain their original settings.

Table 7: Contracts used in this paper.

| Experiment | State requirement | Task requirement |
|---|---|---|
| RoboMNIST / TACO | Backward .1 s velocity, .1 m/s | Displacement at .1 and .2 s, .02 m each |
| Acceleration repair | Same velocity; backward acceleration, .5 m/s$^2$ | Same two endpoints, .02 m each |
| MUKCa | Two recorded configurations, .01 degrees per coordinate | Cross-socket distance 50 mm, tolerance 1 mm |
| CTU | Seven recorded angles, .01 degrees per coordinate | Reflector position in Leica frame, 1 mm |

The robot and object references are the published position/time series. Panda pose repeatability ($<$ $\pm$.1 mm) describes repeated execution (Franka Emika, 2020); it does not provide a per-point bound for the RoboMNIST trajectory. TACO reports marker calibration within 1 mm (Liu et al., 2024, Supplement, Sec. B.1); the audited object origin additionally depends on registration and motion. Its reported cross-modal timestamp mismatch reaches 17 ms (Sec. B.2). We therefore treat the position scans and separate 2 ms time envelope as assumed error sets.

For CTU, each pose includes a reported $U95_i$ from the Leica measurement process (Stepanova et al., 2022). We scale these values into conditional Euclidean reference balls. The .01-degree requirement constrains recorded joint coordinates; it is not an independently calibrated mechanical-angle error bound.

For MUKCa, the socket relations provide the task references. The joint tolerance constrains recorded coordinates. Reference-distance radii of .1 and .2 mm are conditional sensitivity settings, not certified per-record measurement bounds.

For window length $w$, radius-$e$ position errors change a backward velocity by at most $2e/w$, a displacement by $2e$ and the acceleration formed from adjacent velocities by $4e/w^2$. At $w = .1$ s these are $20e$, $2e$ and $400e$ in SI units. The paired repair analysis shares each position error across its derived quantities (Appendix D.2).

# B Geometric Certificates

## B.1 Readout-independent failure

For any fixed scene with $K_Z = \varnothing$, every output fails at least one state or task condition. Thus $\{K_Z = \varnothing\}$ is contained in $\{R(X) \notin K_Z\}$ for every readout and every realization of its randomness. Taking probabilities gives Eq. (4). Partitioning the empty-$K_Z$ event by whether the task intersection is empty gives $\Pr(D)+\Pr(T)$. On T, every output fails at least one task, proving the task-only bound.

## B.2 Constant-velocity geometry

For $P_h(v) = x_0 + hv$, the task set is the ball with center $(y - x_0)/h$ and radius $\tau/h$. Distance from a point to that ball yields Eq. (3). In the analytic example, $y_h = x_0 + hv + ah^2/2$, so a shift $\delta v$ passes exactly when

$$\delta v \in [ah/2 - \tau/h,\ ah/2 + \tau/h].$$

At $a = 1$, $\tau = .05$ and $h \in \{.5, 1\}$, the intervals are $[.15, .35]$ and $[.45, .55]$ m/s. Their separation proves T for the two-query CV head. The typed CA head $x_0 + hv + ah^2/2$ predicts both targets at the reference state.

### B.3 Affine trajectory heads

For $P_h(u) = A_h u + b_h$, define $g_h(u) = \|A_h u + b_h - y_h\|^2 - \tau_h^2$. A task witness supplies an upper bound on $\delta_{\text{req}}$. For nonnegative multipliers $\lambda_h$, weak duality gives

$$L(\lambda) = \inf_u \left[ \|u - s_{\text{ref}}\|^2 + \sum_h \lambda_h g_h(u) \right], \qquad \delta_{\text{req}} \geq \sqrt{\max(0, L(\lambda))}.$$

We independently recompute the saved witnesses with 60-digit scalar arithmetic. Classification requires checked primal points or positive separation margins. On the 97 TACO D windows, the independently checked distance bounds differ by at most $3.7 \times 10^{-11}$ m/s; the smallest necessary-shift lower bound is .10084 m/s. ReLU and kinematic heads use the global enclosures described in Appendices C.4, E and F.

## C RoboMNIST Protocols and Verification

### C.1 Affine training and evaluation

The frozen affine head and readouts use robot 1 at high speed. Head and readout fitting use activities 0–5, repetitions 1–12. Readout selection fits on 0–3, selects on 4–5 and refits on 0–5; activities 6–7 form the old development group. S3 comprises activities 8–9, repetitions 1–12: 24 recordings and 720 windows. The head fits 24 displacement coefficients by recording-balanced least squares on 2,160 training windows, before readout development.

At each anchor, the latest two arrived feedback samples supply a short extrapolation. The input contains two past velocities over .2 s, with no activity, recording or phase identifiers. Anchors start at .2 s and are spaced by .5 s. Eligible windows have finite positions, increasing timestamps, gaps at most .05 s and complete reference support. Replacing future positions with missing values leaves all inputs unchanged in the 72 new recordings; replay of the 24 old development recordings reproduces saved outputs exactly.

**Readouts and selection.** For current and preceding causal velocity estimates $\tilde{s}, \tilde{v}_{\text{prev}}$, the affine readout is

$$u = \tilde{s} + .1\alpha X\Theta, \qquad X = [\tilde{s}^\top/.1, \tilde{v}_{\text{prev}}^\top/.1, 1].$$

Recording-balanced ridge fits the 21 coefficients with $\lambda \in \{10^{-4}, 10^{-2}, 1\}$ and $\alpha \in \{.25, .5, 1\}$. The constant control fits only an intercept. State supervision minimizes velocity squared error normalized by $.1^2$; task training minimizes the mean of two endpoint errors normalized by $.02^2$; the direct joint objective averages all three. Set-target regression fits the closest task-feasible training output, using a minimax point on T and the reference when the teacher search is unresolved.

State selection minimizes validation state MSE; task selection maximizes task passing. Direct joint, constant and set-target selection maximize joint passing. The joint-selected control searches the full non-set library, including past velocity. Passing-rate ties use the other passing rate, state MSE and declared parameter order. It selects the task-trained coefficients at $\alpha = .5$; task selection uses $\alpha = 1$. Both select $\lambda = 10^{-4}$ and refit on activities 0–5. The published-reference row is a privileged scoring control.

Table 8: S3: 24 held-out recordings, 720 windows. Rates are percentages; state RMSE is in m/s.

| Readout | State RMSE | Semantic failure | Task pass | Joint pass | D-task |
|---|---|---|---|---|---|
| Past velocity | 0.0201 | 0.56 | 49.86 | 49.86 | 0/190 |
| State-supervised | 0.0209 | 0.56 | 49.58 | 49.58 | 0/190 |
| Constant correction | 0.0201 | 0.56 | 49.86 | 49.86 | 0/190 |
| Joint-selected | 0.0484 | 3.19 | 56.11 | 55.42 | 1/190 |
| Direct joint objective | 0.0463 | 2.64 | 55.69 | 55.00 | 1/190 |
| Task-trained | 0.0918 | 25.28 | 64.17 | 52.50 | 18/190 |
| Set-target | 0.0488 | 3.61 | 54.17 | 53.47 | 1/190 |
| Published reference | 0.0000 | 0.00 | 50.28 | 50.28 | 0/190 |

The 18 necessary-shift lower bounds range from .10655 to .21774 m/s; actual readout shifts range from .11285 to .25306 m/s, using the same backward-window reference.

**Primary comparison and uncertainty.** On every successful S3 D window, task passing and state deviation above .1 m/s are checked separately from the necessary-shift lower bound. Pooled D-task success is 18/190 (9.47%). Averaging within recordings and weighting activities equally gives 9.73%. A paired bootstrap of recordings within each activity, 10,000 samples with seed 20260909, gives a 95% interval of $[6.71, 12.71]$ points above state supervision, conditional on these activities and frozen models.

Over all S3 windows, task gain is 14.58 points $[11.25, 17.92]$, with 23/24 recordings improving; joint gain is 2.92 points $[-.14, 6.11]$. Joint-selected passing is 55.42%, versus 52.50% for task training. Figure 3 gives all 18 necessary bounds and actual shifts, with record identities and margins in the supplementary results.

### C.2 Reference robustness of the realized D successes

We hold the affine head and all 18 successful learned outputs fixed. Independent radius-$e$ position balls give velocity and endpoint radii $20e$ and $2e$. An actual D success remains certified when the saved dual excludes the enlarged semantic ball, a task witness survives tightened endpoint radii, and the learned output satisfies $\max_h \|P_h(\hat{u}) - y_h\| + 2e \leq .02$ m.

Table 9: Reference robustness among the same 18 nominal affine D successes. The final column requires both D stability and task success of the saved output.

| Assumed reference envelope | D stable | Actual D success certified |
|---|---|---|
| Nominal | 18 | 18 |
| 1 mm position | 9 | 4 |
| 1 mm position + 2 ms time | 5 | 0 |

At 1 mm, exclusion, task-witness robustness and learned task success separately retain 9/18/11 windows; their intersection contains four, in different activity-9 repetitions. Supported joint position radii range from .193 to 2.178 mm. The time envelope uses $e = .001 + .002L$ with saved local slope $L$. The per-case results give component radii and curves, including the maximum 1.384 ms additional time radius at 1 mm position. Missing sufficient certificates leave robustness unresolved. Appendix D.2 extends the check to actual joint success after an interface change.

### C.3 Nonlinear RoboMNIST heads: fitting and frozen tests

The state and endpoint radii remain .1 m/s and .02 m at horizons .1/.2 s. The head is $P(u) = Lu + .02g_\theta(u/.1)$, where $L = [.1I_3; .2I_3]$ and $g_\theta$ is a $3 \to w \to 6$ ReLU MLP with zero-initialized final layer. Four candidates combine widths 16/32 with weight decay $10^{-4}/.01$. Full-batch Adam minimizes normalized endpoint MSE on activities 0–3, repetitions 1–12, using learning rate .003, at most 1,200 epochs and checks every 20 epochs. Reference-input MSE on activities 4–5 selects width 32, weight decay .01 and 1,160 epochs using seed 40101 (development MSE $8.94037 \times 10^{-5}$ m$^2$). The selected duration is refitted on activities 0–5 for seeds 40101–40103.

For the same six past features $x$, each readout is $u = \tilde{s} + .1f_\phi(x)$, a $6 \to 32 \to 3$ ReLU MLP with zero-initialized final layer. With $\ell_s = \|u - s\|^2/.1^2$ and $\ell_h = \|P_h(u) - y_h\|^2/.02^2$, the state, task and joint objectives are $\ell_s$, $(\ell_{.1} + \ell_{.2})/2$ and $(\ell_s + \ell_{.1} + \ell_{.2})/3$. Adam uses learning rate .002, weight decay $10^{-4}$, at most 1,000 epochs and checks every 20 epochs. Validation selects state RMSE or the corresponding passing rate, with ties resolved by the other rate, state RMSE and earliest epoch. Selected durations are refitted on 0–5.

Both affine and nonlinear heads receive all three newly trained MLP objectives. The original affine readouts stay frozen as separate controls. The original affine development scores use its already fitted 0–5 head; nonlinear readout development uses a head fitted on 0–3. Final heads exclude activities 8–9. Training uses CPU float64, PyTorch 2.6.0, NumPy 1.26.4 and SciPy 1.14.0.

S3 is reused retrospectively. Before accessing the new trajectories we freeze all models and three transfer groups: robot 1 at medium/low speed and robot 2 at high speed, activities 8–9, repetitions 1–12. The active robot in the filename selects its position entry; both robots are instances of the same type. Published coordinates are retained without adaptation. All 72 new recordings pass the affine eligibility checks, giving 30 windows each. Missing future positions leave every input unchanged; the 18 frozen MLPs replay S3 within $8.89 \times 10^{-16}$ m/s.

### C.4 Partial certification for the ReLU heads

Verification of piecewise-linear networks is an established problem (Tjeng et al., 2019). Here we use bounded interval and affine-envelope checks to audit the physical constraints, retaining the original Euclidean balls.

**Shared labels and task witnesses.** For a given head seed and window, the candidate pool contains outputs of every evaluated readout under that head, plus past velocity and the reference. A verified candidate inside both task sets and the semantic ball establishes F. If a candidate passes the tasks but no joint candidate is found, we attempt to exclude all of $B_{\cdot 1}(s)$. Complete exclusion establishes D; a found joint point establishes F. A missing task witness or incomplete exclusion leaves U, since failed search leaves feasibility unresolved. All compared readouts share the resulting label vector. The physical definition of D is independent of the readout. The subset we certify depends on the pooled witnesses and finite verification budget, so D-success counts describe that subset rather than population prevalence.

**Exported mathematical model.** The stored effective parameters define

$$P(u) = Lu + c + V\,\mathrm{ReLU}(Wu + b),$$

where the input and output scales above are absorbed into $W, V, c$. The six outputs retain their original physical units. All exported parameters are round-tripped float64 values; the independent checker interprets each as its exact binary rational value. On S3 reference inputs, head outputs were checked against the training implementation with relative and absolute tolerances $10^{-12}$. The final certificates use the stored readout outputs and $s, y$ derived from the published references. They apply to this mathematical model; floating-point execution on a particular hardware implementation requires a separate analysis.

**Excluding a semantic ball.** The initial box $[s - \epsilon, s + \epsilon]$ covers the full closed Euclidean ball. A box can be discarded when its distance from $s$ exceeds $\epsilon$. Otherwise interval propagation encloses each preactivation by $[z_l, z_u]$ and each endpoint output coordinate by an interval. If the resulting lower bound on an endpoint's Euclidean error exceeds $\tau$, that box contains no jointly passing output.

For tighter bounds, choose any finite direction $q$ for one endpoint. Active and inactive ReLUs have exact affine representations. On an interval crossing zero, a nonnegative coefficient can use either 0 or $z$ as a lower ReLU envelope. A negative coefficient uses the upper chord $\mathrm{ReLU}(z) \le z_u(z - z_l)/(z_u - z_l)$, whose inequality reverses after multiplication. These choices produce an enclosing affine lower form $q^\top(P_h(u) - y_h) \ge a^\top u + d$ throughout the box. We bound its minimum on the box and on the full ball; their maximum remains a lower bound on the intersection. For exact coefficients the ball bound is

$$a^\top s - \epsilon\|a\|_2 + d.$$

Exclusion requires this lower support, or the box bound, to exceed an upper bound on $\tau\|q\|_2$. Thus no approximate unit-norm assumption is needed. Interval coefficient enclosures may weaken the bound but cannot justify a stronger exclusion.

Every subdivision covers its parent with two closed boxes. Only exhaustion of all boxes, together with a separately verified task witness, yields D. The float screening pass uses a $10^{-7}$ m guard and at most 20,000 nodes or 1.5 seconds per window. It includes a search for joint points within retained boxes. Timeouts remain U.

**Independent checks and margins.** A separately implemented checker converts stored floats exactly and uses 80-digit Decimal interval arithmetic with downward and upward rounding. Square roots are expanded to adjacent representable bounds. It uses a cyclic $y, z, x$ split, distinct from the

float screen's longest-side split, and a further $10^{-30}$ guard. Each original D check has a 60,000-node/60-second budget. All 86 head-seed/window D certificates pass this check. The 7,060 F witnesses, all D task witnesses and all 64 task-trained D realizations also pass separate directed-interval point checks. These totals count seed–window pairs: the 86 certificates and 64 realizations can share physical scenes and trajectories. Table 10 reports the original F/D/U denominators and verification time.

For an actual D success, the full-ball exclusion establishes $\delta_{\text{req}} > .1$ m/s. Its observed shift is a separate quantity, with a positive lower margin above .1 m/s; both endpoint errors have upper bounds below .02 m. The per-window results include these task and state margins and the checker node counts. These nonlinear certificates use nominal references. The affine reference-error analysis applies only to its own head and saved outputs.

Table 10: Shared partial labels and actual realizations, separately for every nonlinear head. Screening time sums the recorded float checks; verification time sums the independent Decimal D checks. Times exclude training and data preparation. All D checks succeed.

| Setting | Seed | F | D | U | D-task | Screen (s) | Verify (s) |
|---|---|---|---|---|---|---|---|
| R1 high | 40101 | 546 | 16 | 158 | 12 | 49.31 | 4.90 |
| R1 high | 40102 | 544 | 10 | 166 | 7 | 52.37 | 4.20 |
| R1 high | 40103 | 548 | 9 | 163 | 2 | 38.13 | 19.09 |
| R1 medium | 40101 | 592 | 3 | 125 | 3 | 22.46 | 0.04 |
| R1 medium | 40102 | 588 | 3 | 129 | 3 | 26.70 | 0.03 |
| R1 medium | 40103 | 588 | 1 | 131 | 1 | 13.23 | 0.01 |
| R1 low | 40101 | 657 | 2 | 61 | 2 | 1.43 | 0.08 |
| R1 low | 40102 | 654 | 4 | 62 | 3 | 5.97 | 0.09 |
| R1 low | 40103 | 648 | 2 | 70 | 1 | 1.52 | 0.08 |
| R2 high | 40101 | 564 | 16 | 140 | 15 | 41.83 | 13.90 |
| R2 high | 40102 | 563 | 12 | 145 | 10 | 44.47 | 1.68 |
| R2 high | 40103 | 568 | 8 | 144 | 5 | 28.18 | 1.08 |

### C.5 Nonlinear results, controls and uncertainty

Table 3 reports all three seeds' D-task successes and mean performance by objective. Table 10 retains the per-seed F/D/U denominators. The complete head/readout scores, including original affine baselines, past velocity and the privileged reference, are included in the supplementary results.

On S3, the affine partition is 527 F, 190 D and three T. The nonlinear heads certify 546/544/548 F, yielding lower bounds of 19/17/21 additional feasible windows; unresolved cases remain in the denominator. Their 21 seed-window D successes cover 15 distinct windows in 12 recordings. The medium-, low-speed and robot-2 groups cover 3/3/17 successful windows in 3/3/10 recordings. Seeds share data and can succeed on the same window.

For intervals, each record's score is first averaged over the three fitted seeds. We resample whole records within each activity 10,000 times (seed 40601), conditional on the fitted models and the two activities. At low speed, task minus state training has mean task gain $-.83$ points with interval $[-2.45, .69]$. Joint training has the highest mean joint passing among the three MLP objectives in every setting. Replacing the affine head and retraining that matched objective improves joint passing by 4.54–13.89 points. This comparison uses a different training protocol from the original affine joint-selected baseline.

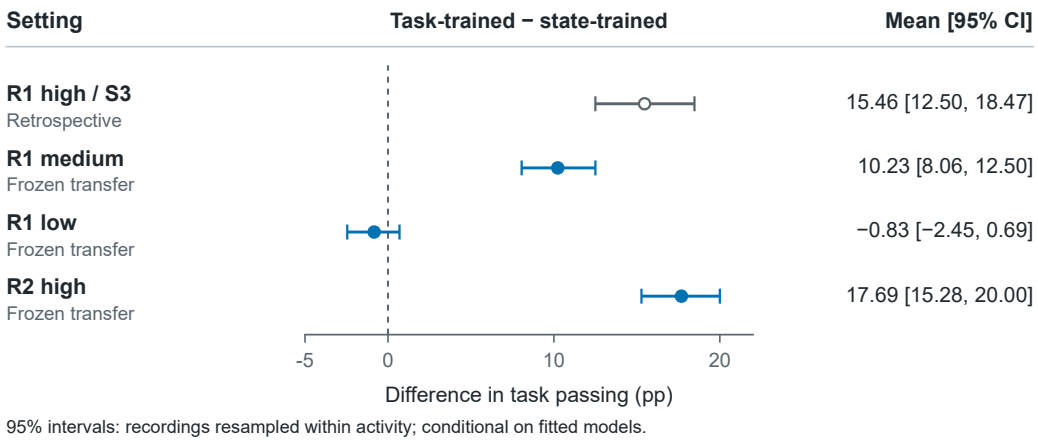


Figure 5: Paired task-passing differences under the same nonlinear head, averaged over three fitted seeds. Bars are the saved conditional 95% intervals from resampling recordings within activity, conditional on these models and two activities. R1 high/S3 is retrospective; R1 medium, R1 low and R2 high test frozen models on newly accessed recordings. All four settings are shown; the low-speed interval crosses zero. Appendix C.5 describes the resampling; full scores and paired counts are included in the supplementary results.

## D Acceleration-Interface Repair

**Protocol and state.** Training uses R1/high activities 0–5, repetitions 1–12; development uses activities 6–7, repetitions 1–12. Selected durations are refitted on activities 0–7. Confirmation uses activities 8–9, repetitions 25–30, in all four settings. Each of the 48 records has 30 eligible windows. The position reference and past-input construction follow Appendix C.1; replacing future positions by missing values leaves every input unchanged. The six-dimensional state consists of backward .1-second velocity and the difference between two adjacent backward velocities divided by .1 s. The evaluation protocol and pre-evaluation hash fix the .5 m/s$^2$ acceleration radius and .25/.5/1.0 m/s$^2$ sensitivity grid before confirmation. The radius is a design requirement: $h(h+w)\epsilon_a/2 = 15$ mm at $h = .2$ s and $w = .1$ s. Table 7 distinguishes it from reference uncertainty.

**Heads, readouts and comparison identity.** The five original candidates are the frozen affine and nonlinear heads, ordinary CA, window-matched CA and window-matched CA with a bounded residual. The residual takes the six state coordinates normalized by (.1 m/s, .5 m/s$^2$), uses a $6 \to 32 \to 6$ ReLU MLP, and applies .005 tanh to its output in meters. A $6 \to 32 \to 6$ readout adds a scaled correction to the causal state estimate. All candidates share this readout architecture and full scoring contract; the velocity-only heads ignore the last three inputs. These new readouts differ from the $6 \to 32 \to 3$ models in Table 3.

Adam trains readouts for at most 1,000 epochs, with learning rate .002, weight decay $10^{-4}$ and development checks every 20 epochs. State, task and joint objectives average, respectively, the two normalized state-group losses, the two endpoint losses, or all four. State models select by normalized state RMSE, task models by task passing, and joint models by full joint passing, followed by the declared secondary scores and earliest epoch. Residual heads select by reference-input MSE using at most 1,200 epochs, learning rate .003 and weight decay .01. Seeds 40101–40103 are all retained.

The additional data/budget-matched nonlinear control uses the old $3 \to 32 \to 6$ head architecture with the same training/development records, 1,200-epoch budget and reference-MSE selection as the new residual head, then refits on activities 0–7. This control was added after the first confirmation results to remove the original nonlinear head's smaller training set as a confound. The same 48 confirmation records, thresholds and seeds are retained.

**Paired certification and actual success.** The new semantic set is the product of the velocity and acceleration balls. F requires a verified state and task witness. D combines a task witness with global

semantic exclusion; incomplete searches remain U. For affine and CA heads, nonnegative support bounds account for both balls. The residual exclusion subtracts the worst componentwise 5 mm residual contribution from the CA support bound. ReLU heads use the guarded exclusion procedure. Independently rechecked primal points, 60-digit support calculations and full ReLU exclusion replay validate the saved labels. Affine T labels are separately checked by a nonnegative weighted squared-error dual. These new checks provide guarded numerical verification; the directed-interval procedure for the older experiment is described separately in Appendix C.4.

Physically identical affine and CA heads share a pooled witness label across seeds. Learned heads retain separate labels. The paired old affine partitions are 262/96/2/0, 307/52/0/1, 337/22/0/1 and 267/93/0/0 (F/D/T/U), in table order. For each saved joint-trained output, we independently recompute both state errors and both head endpoint errors, matching all saved scores. We then count

$$n_{D\to J} = \sum_i \mathbf{1}\{L_i^{\text{affine}} = D\}\mathbf{1}\{\hat{z}_i^{\text{residual}} \in K_i^{\text{residual}}\}.$$

Every counted output is also a verified F witness, while old D windows have zero joint passing under the affine head. The unions across seeds cover 65, 30, 11 and 55 distinct physical windows in 12, 12, 9 and 12 recordings. Table 11 retains the per-seed counts and the effects on old F windows.

Table 11: Paired repair on 360 windows per setting, with old affine labels. Triples list seeds 40101/40102/40103; $F \to D$ is identical across seeds. $D \to J$ counts actual joint success, and U denotes unresolved windows.

| Setting | Old D | $D \to F$ | $D \to J$ | $F \to D$ | $F \to U$ |
|---|---|---|---|---|---|
| R1 high | 96 | 78/81/84 | 60/52/60 | 8 | 8/12/10 |
| R1 medium | 52 | 44/45/44 | 29/28/29 | 3 | 11/6/4 |
| R1 low | 22 | 14/14/14 | 11/11/11 | 3 | 12/8/11 |
| R2 high | 93 | 81/78/80 | 52/51/52 | 3 | 10/16/11 |

**Uncertainty and model selection.** Record-level scores are first averaged over the three fixed seeds. Within each of activities 8 and 9, six whole recordings are resampled with replacement, for 5,000 paired bootstrap samples. Intervals are conditional on the fitted models and these activities and are pointwise, without multiplicity correction. Residual minus matched nonlinear gains are 5.65 [2.41,8.52], 3.15 [1.11,5.09], 1.85 [−.28,4.17] and 4.91 [3.06,6.76] points. Differences use unrounded record scores. The numerical supplement retains complete intervals for all controls and objectives. The following decomposition and reference scans are retrospective analyses of the same frozen outputs.

Table 12: Full joint passing (%) for joint-trained readouts. Cells give three-seed means at acceleration radii .25/.5/1.0 m/s$^2$. Velocity and task tolerances stay fixed; all six heads are shown.

| Head | R1 high | R1 medium | R1 low | R2 high |
|---|---|---|---|---|
| Original affine | 50.65/52.22/52.31 | 57.50/60.28/60.56 | 65.93/70.19/71.20 | 54.72/55.93/55.93 |
| Frozen nonlinear | 64.44/65.56/65.65 | 68.89/71.20/71.48 | 70.65/75.37/76.02 | 67.04/67.78/67.87 |
| Data/budget-matched nonlinear | 63.98/65.83/66.11 | 67.59/70.00/70.37 | 71.11/75.65/76.30 | 66.76/67.69/67.69 |
| Ordinary CA | 61.85/68.70/68.89 | 66.02/73.43/73.61 | 63.43/76.39/77.04 | 62.13/70.09/70.19 |
| Window-matched CA | 58.52/69.17/70.19 | 60.83/73.06/74.26 | 60.83/74.91/76.11 | 57.78/70.56/72.59 |
| Window-matched CA + residual | 60.74/71.48/72.04 | 62.78/73.15/73.98 | 64.63/77.50/78.43 | 61.20/72.59/73.89 |

### D.1 Acceleration requirements and joint passing

Let $C$ be the fraction of outputs passing the velocity and both endpoint requirements, and $A_{\epsilon_a}$ the fraction passing these three checks but failing acceleration. Full joint passing is $J_{\epsilon_a} = C - A_{\epsilon_a}$. This decomposition holds for every head, seed and window in Table 12.

Table 13: Residual minus data/budget-matched nonlinear, in percentage points. At acceleration radius .25 m/s$^2$, the extra acceleration exclusions exceed the gain in velocity-and-endpoint passing. All entries average the three fixed seeds.

| Setting | $\Delta C$ | $\Delta A_{.25}$ | $\Delta J_{.25}$ |
|---|---|---|---|
| R1 high | +5.93 | +9.17 | -3.24 |
| R1 medium | +3.61 | +8.43 | -4.81 |
| R1 low | +1.94 | +8.43 | -6.48 |
| R2 high | +6.20 | +11.76 | -5.56 |

At the default .5 m/s$^2$ radius, the extra exclusions are .28, .46, .09 and 1.30 points, respectively, leaving a positive net gain. At .25, they reverse the ranking in all settings. Passing near the threshold also differs from average error: R1 low has lower acceleration RMSE under the residual head (.3475 versus .3571 m/s$^2$), despite its lower passing rate at .25. The numerical supplement retains all component errors, paired recording intervals and the 16 combinations of failed requirements.

### D.2 Old D and new joint success under reference uncertainty

We require old affine D and actual residual joint success for every reference in a shared uncertainty set. Each interpolated position at $t-.2, t-.1, t, t+.1, t+.2$ s lies in an independent Euclidean radius-$e$ ball. Shared positions carry the same errors into each derived quantity; heads and saved outputs stay fixed.

**Actual output.** Write $E_v, E_a, E_1, E_2$ for its nominal velocity, acceleration and endpoint errors. With $w = .1$ s, the largest changes in those references are $20e$, $400e$, $2e$, $2e$. Thus the saved output is jointly successful throughout the position-error set exactly when

$$E_v + 20e \le .1, \qquad E_a + 400e \le \epsilon_a, \qquad E_h + 2e \le .02 \quad (h = 1, 2). \tag{6}$$

Each bound is attainable by choosing the contributing position errors along opposite directions. Although the maximizing errors can differ between constraints, failure of any one constraint suffices to break joint success.

**Old interface.** For saved affine support vectors $q_h$, set $c = \sum_h A_h^\top q_h$ and

$$m = \sum_h q_h^\top (A_h s_w + b_h - y_h) - .1\|c\| - .02 \sum_h \|q_h\|.$$

Expanding the shared position errors gives the exact worst decrease $eL$ for this fixed support direction, where

$$L = \|c/w + q_1 + q_2\| + \|c/w\| + \|q_1\| + \|q_2\|. \tag{7}$$

We require $m - eL > 10^{-9}$ m for semantic exclusion and a saved affine task witness whose endpoint errors plus $2e$ stay within .02 m. The affine head ignores acceleration, so its exclusion applies to every acceleration ball in the scan. We reuse saved support directions and witnesses, taking the strongest valid fixed evidence without fitting a new output. Support margins are recomputed with 60-digit arithmetic and independently checked by constructing the worst position perturbations.

**Coverage.** All four settings retain paired evidence at .1 and .5 mm (Table 14). At 1 mm, R1 low retains no paired certificate; at 2 mm all settings reach zero. Indeed, $e > 1.25$ mm makes the acceleration condition impossible at the default radius, even for zero nominal acceleration error. Losing the old sufficient certificate alone leaves compatibility unresolved. These scans assume position-error sets; they do not calibrate the published trajectories or perturb the estimator's inputs.

The overall ranking can change while paired repair cases persist. Residual minus data/budget-matched nonlinear worst-case joint passing is $+3.15, -.37, -1.02, -.46$ points at .5 mm and $-5.83, -8.33, -16.02, -8.06$ at 1 mm, in table order. The numerical supplement retains every position/acceleration combination and separate exclusion and witness counts. These position-only tests make no timing-error claim.

Table 14: Paired old D and actual new joint success certified for every allowed reference, at $\epsilon_a = .5$ m/s$^2$. Each entry lists seeds 40101/40102/40103. These seeds share the same 360 windows per setting; their counts are not independent cases.

| Position radius (mm) | R1 high | R1 medium | R1 low | R2 high |
|---|---|---|---|---|
| 0 | 60/52/60 | 29/28/29 | 11/11/11 | 52/51/52 |
| 0.05 | 58/51/58 | 28/27/28 | 9/9/9 | 51/50/51 |
| 0.1 | 56/50/56 | 28/27/28 | 9/9/9 | 49/49/49 |
| 0.25 | 48/48/53 | 24/25/25 | 9/9/9 | 45/48/45 |
| 0.5 | 41/43/41 | 18/18/17 | 4/4/4 | 39/37/38 |
| 1 | 8/5/5 | 4/3/5 | 0/0/0 | 8/8/8 |
| 2 | 0/0/0 | 0/0/0 | 0/0/0 | 0/0/0 |

# E MUKCa: Relational Certificates and Recovery

**Models, records and relations.** The frozen source is Franzese et al. (2026)'s release, commit `58cb04a203d3`. Eligibility requires both URDFs, fitting KPIs and raw records for both sockets. The six- or seven-angle chain ends at `ball_link`; off-chain gripper joints are excluded. No model is refitted. Fitting KPIs list `front` except for Panda 2's `frontcenter`. Table 17 lists the other locations; this split follows published metadata.

All numeric rows, including the first, are retained. We omit the source reader's first-row skip and model-dependent outlier filter. Published KPI positions match raw-row predictions within $8 \times 10^{-16}$ m. Each pair is audited separately, using all distinct unordered same-socket pairs ($t = 0$) or all cross-socket pairs ($t = .05$ m). Shared records are not independent samples. The state box constrains both configurations, with angles converted to radians; primary requirements and the distinction from measurement uncertainty are in Table 7.

**Global distance enclosure.** For active joint $k$, let $L_k$ be the sum of the lengths of all fixed URDF translations after that joint, including the terminal tool. Each term is multiplied by rotations of norm one. Unit-axis rotation derivatives have operator norm at most one, giving global bounds

$$\|\partial_k f\|_2 \le L_k, \qquad \|\partial_k \partial_l f\|_2 \le L_{\max(k,l)}, \qquad H = \sum_{k,l} L_{\max(k,l)}.$$

The index $\max(k, l)$ denotes the later joint in the chain. For pair displacement $v = f(q_i) - f(q_j)$ with $d = \|v\|_2 > 0$, set $n = v/d$, $A = \|J_i^\top n\|_1 + \|J_j^\top n\|_1$ and $r = 2\epsilon_q \sum_k L_k$. The total two-state Taylor remainder is at most $\epsilon_q^2 H$. Projection onto $n$ and the global displacement radius give

$$\ell = \max\{0, d - \epsilon_q A - \epsilon_q^2 H\}, \tag{8}$$

$$u = \min\left\{ d + r,\ d + \epsilon_q A + \epsilon_q^2 H + \frac{r^2}{2(d-r)} \right\} \quad (d > r). \tag{9}$$

For $d \le r$, use $u = d + r$; when $d = 0$, use $\ell = 0$. The second upper bound follows because the segment of displacement vectors stays at distance at least $d - r$ from zero, where the norm Hessian is bounded by $1/(d - r)$. Thus every allowed distance lies in $[\ell, u]$. Disjointness from $[t - \tau, t + \tau]$, with a $10^{-9}$ m exclusion guard, rules out the entire semantic box. Forty bisections over 0–.1 radians retain the largest verified excluded radius as a lower bound on $\delta_{\text{req}}$. Loss of exclusion at the other endpoint does not certify attainability.

**Witnesses and independent checks.** Same-socket task feasibility follows by setting the two output configurations equal. For cross-socket feasibility, each model has a saved pair whose distance is checked directly within .05 mm of 50 mm. A local numerical search supplies that pair when needed; all saved pairs satisfy the listed URDF joint limits. These mathematical task witnesses establish per-pair feasibility. Recovery uses the calibrated model's output at the original recorded pair.

Independent elementary-transform FK and enclosure calculations reproduce all 19 locations and 1,824 tolerance cells. The largest FK difference is $7.8 \times 10^{-16}$ m. Analytic Jacobians agree with central differences, and 4,560 sampled pair perturbations satisfy the enclosures. The global derivative

bounds justify full-box coverage; the numerical checks test their implementation. Per-pair results include record identifiers, model hashes, witnesses and bounds.

### E.1 Same-socket results and sensitivity

Table 15 gives the other physical relation under the primary contract. Passing-pair counts increase at all 14 locations. The same-socket task alone admits coincident outputs, so the main comparison also requires the nonzero cross-socket separation.

Table 15: Same-socket relations on the same 14 locations and 838 records. Lost counts previously passing pairs that fail under the calibrated model.

| System | Pairs | Old→new pass | Recovered / certified D | Lost |
|---|---|---|---|---|
| FR3 | 2,387 | 6→2,214 | 2,206/2,377 | 1 |
| Panda 2 | 4,936 | 32→4,935 | 4,839/4,840 | 0 |
| Panda 3 | 2,344 | 16→2,261 | 2,223/2,306 | 0 |
| KUKA | 1,686 | 829→1,586 | 365/425 | 16 |
| Kinova | 1,221 | 19→64 | 51/1,187 | 13 |

The frozen grid uses joint radii .001/.005/.01/.02 degrees, distance tolerances .25/.5/1/2 mm and assumed relation-error radii $\eta = 0, .1, .2$ mm. For the latter, old exclusion uses $[t-\tau-\eta, t+\tau+\eta]$; the old task witness and calibrated record distance must each satisfy error$+\eta \leq \tau$ (with the guard also applied to the witness). These are conditional reference-error checks. MUKCa's small-uncertainty assumptions do not supply a certified per-record radius.

Table 16 reports recovered D over certified D as one setting is changed at a time. The D set can change between columns. An entry of 0/0 means no D was certified, so a recovery rate is undefined. All combinations and locations, including the fitting locations, are included in the supplementary results.

Table 16: Cross-socket recovery when one requirement changes. Entries are recovered / certified D. Other values stay at .01 degrees, 1 mm and $\eta = 0$. The last column assumes a .2 mm relation-reference error radius.

| System | $\tau = .25$ mm | $\tau = .5$ mm | $\tau = 2$ mm | $\epsilon_q = .02^\circ$ | $\eta = .2$ mm |
|---|---|---|---|---|---|
| FR3 | 1,171/2,280 | 1,832/2,234 | 1,909/1,909 | 1,953/1,986 | 1,991/2,071 |
| Panda 2 | 3,202/4,333 | 3,999/4,115 | 2,893/2,893 | 3,195/3,195 | 3,558/3,558 |
| Panda 3 | 1,198/2,227 | 1,871/2,184 | 1,952/1,952 | 2,019/2,019 | 2,063/2,084 |
| KUKA | 387/433 | 320/320 | 0/0 | 1/1 | 21/21 |
| Kinova | 71/1,167 | 146/1,120 | 328/798 | 188/958 | 161/974 |

### E.2 Location-level counts

Table 17 sums to Table 6 within each system. The complete location-level results cover all 19 locations and both relation types.

## F CTU Motoman: external kinematic interface

**Source and fit.** The standard-DH right-arm and reflector model follows Stepanova et al. (2022)'s release, commit `a20f0454732b`. Following `loadNewLeicaDataset.m`, we retain one finite Leica row per pose and remove retained row 13 (pose 13), leaving 586 poses. IDs divisible by five give the fixed 121-pose holdout; the other 465 fit registration and calibration. Both sets come from one acquisition.

The original head uses the released DH parameters and an SVD rigid fit. Calibration minimizes development position residuals over the whitelist in `loadMotomanLeica.m` and six registration

Table 17: Cross-socket results at all 14 locations in the main summary, at the primary .01-degree / 1 mm requirements. Records are the two sockets' raw rows; each relation pairs one record from each socket. Recovered counts old D pairs passing at unchanged inputs. Lost counts previously passing pairs that now fail.

| System / location | Records | Relations | Certified D | Recovered | Lost |
|---|---|---|---|---|---|
| FR3 / right | 69 | 1,184 | 989 | 960 | 1 |
| FR3 / very-front | 71 | 1,260 | 1,128 | 1,120 | 1 |
| Panda 2 / backcenter | 66 | 1,089 | 796 | 796 | 0 |
| Panda 2 / backleft | 63 | 992 | 725 | 725 | 0 |
| Panda 2 / backright | 66 | 1,088 | 811 | 811 | 0 |
| Panda 2 / frontleft | 64 | 1,024 | 734 | 734 | 0 |
| Panda 2 / frontright | 60 | 900 | 669 | 669 | 0 |
| Panda 3 / left | 60 | 900 | 762 | 762 | 0 |
| Panda 3 / right | 77 | 1,452 | 1,345 | 1,345 | 0 |
| KUKA / left_back | 62 | 961 | 69 | 69 | 0 |
| KUKA / raised_left | 56 | 784 | 0 | 0 | 0 |
| Kinova / front_rotated | 40 | 400 | 329 | 38 | 31 |
| Kinova / left | 41 | 420 | 345 | 109 | 31 |
| Kinova / right-high | 43 | 462 | 345 | 64 | 45 |

variables with the bounds below; least squares converges in six evaluations. An independent quaternion solution for registration alone matches the SVD predictions within $2.89 \times 10^{-15}$ m. Its holdout mean error stays at 2.550 mm and passing at 7/121; calibration gives 2.096 mm and 16/121.

Free DH entries are $(a, \alpha, \theta_0)$ in row 3, $(a, d, \alpha, \theta_0)$ in rows 4–6 and $(a, d, \theta_0)$ in reflector row 8. Length, twist and joint-offset changes are bounded by .05 m, .05 rad and .1 rad, respectively. Registration rotation-vector and translation increments have coordinate bounds .2 rad and .2 m. All other entries stay fixed.

Table 18: CTU transitions at the unchanged recorded inputs. The six losses are losses of reference-point passing; they do not establish new D classifications.

| Original model | Calibrated pass | Calibrated fail | Total |
|---|---|---|---|
| Recorded input passes | 1 | 6 | 7 |
| Recorded input fails, certified D | **12** | 87 | 99 |
| Other recorded-input failures | 3 | 12 | 15 |
| Total | 16 | 105 | 121 |

The semantic box constrains the turntable and six arm angles to .01 degrees per recorded coordinate. The task radius is 1 mm in the Leica frame. An inverse-kinematics task witness may change each coordinate by at most .1 radians; it is a mathematical input, not an executed robot command. The port is checked against independent elementary transforms.

**Full-box exclusion.** Let $f(q)$ denote the original head in the Leica frame, $e_i = \|f(q_i) - y_i\|_2$, $n_i = (f(q_i) - y_i)/e_i$ and $J_i = \partial f/\partial q(q_i)$. For DH link lengths $\ell_k = \sqrt{a_k^2 + d_k^2}$, including the reflector, put $b_j = \sum_{k=j}^{8} \ell_k$ and $H = \sum_{j=1}^{7} \sum_{k=1}^{7} b_{\max(j,k)}$. The global rotation-derivative argument of Appendix E bounds each mixed derivative by $b_{\max(j,k)}$. The full-box bound is

$$\|f(q) - y_i\|_2 \geq L_i := e_i - \epsilon_q \|J_i^\top n_i\|_1 - \tfrac{1}{2}\epsilon_q^2 H, \quad \|q - q_i\|_\infty \leq \epsilon_q. \tag{10}$$

Here $H = 53.7701$ m includes all 49 mixed terms and the .24910 m reflector displacement. The Jacobian agrees with central differences within $6.09 \times 10^{-10}$ m/rad. All derivative calculations use radians.

For a saved task witness $w_i$ and a reference ball of radius $\rho_i$, we check

$$L_i - \rho_i - \tau > 0, \qquad \tau - \|f(w_i) - y_i\|_2 - \rho_i \geq 0, \qquad \tau - \|f_{\text{cal}}(q_i) - y_i\|_2 - \rho_i \geq 0. \tag{11}$$

The first two inequalities use a $10^{-9}$ m guard and retain old D throughout the reference ball; the third certifies new input success. At $\rho_i = 0, U95_i, 3U95_i$, old D counts are 99/99/99 and recovered counts

12/12/11. These scaled uncertainty balls are conditional, without deterministic or simultaneous 95% coverage. The per-pose results include all margins.

**Verifier refinement and tolerance grid.** The Taylor refinement reuses the pilot's fixed models, poses, witnesses and tolerances. Table 19 gives the five joint radii at 1 mm. The supplementary results include the initial verifier, full grid, development counts and per-pose margins.

Table 19: CTU joint-radius sensitivity, all 121 holdout poses at the 1 mm task requirement. Recovered counts require all three paired conditions. The calibrated recorded input passes on 16 poses at every joint radius.

| Joint radius | Nominal reference | | $3U95_i$ ball | |
|---|---|---|---|---|
| (degrees) | Old D | Recovered | Old D | Recovered |
| 0.001 | 114 | 15 | 113 | 14 |
| 0.005 | 106 | 14 | 105 | 13 |
| 0.01 | 99 | 12 | 99 | 11 |
| 0.02 | 78 | 7 | 76 | 7 |
| 0.05 | 21 | 0 | 20 | 0 |

## G TACO: Prediction Error and Feasibility

The 24-parameter affine head fits 25 official training recordings of the dust/roller/pan interaction (Liu et al., 2024). Leave-tool-ID-out validation selects ridge strength .01 from $0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100$ by recording-balanced MSE. The head is frozen before accessing the 12 official test recordings; definitions follow Table 7.

The test has 400 object windows from one interaction composition. Recording-balanced normalized MSE falls from 1.02603 to .56764 (44.68%). CV has F/D/T counts 300/88/12 and the affine head 292/97/11. Table 20 shows the paired changes: all 32 lost F windows become D, while 23 old D windows and one old T window become F. All twelve recordings have lower MSE; joint feasibility falls in seven, rises in four and is unchanged in one. The supplementary results include the per-record data and plot.

Table 20: Paired CV-to-affine categories on the same 400 TACO windows from twelve recordings. Every category is resolved.

| CV \ affine | F | D | T | Total |
|---|---|---|---|---|
| F | 268 | 32 | 0 | 300 |
| D | 23 | 59 | 6 | 88 |
| T | 1 | 6 | 5 | 12 |
| Total | 292 | 97 | 11 | 400 |

The recording-equal feasibility change is $-1.54$ points, with paired-record 95% interval $[-5.00, 1.94]$. We resample twelve whole records 5,000 times (seed 370909), keeping both heads and their windows together. The exploratory interval assumes exchangeable records; shared capture dates and one interaction composition limit inference beyond this cohort.

Across eight state radii and five endpoint radii, 32 recording-equal changes are negative and eight positive. All 32,000 classifications have independently checked certificates and none is unresolved. The full grid, per-record results and position/time reference scans are included in the supplementary results.