PROBE2ACT: KNOWING WHEN TO STOP PROBING ARTICULATED OBJECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A robot can discover how an unfamiliar drawer or hinged door moves by probing it, but every additional interaction consumes budget. We study a practical decision: when does the observed motion justify stopping identification and beginning execution? Probe2Act combines an analytic common-path estimator with a calibrated stopping rule. The estimator retains displacement between probes, so individually weak responses can jointly reveal a local motion direction; the rule stops when the direction's calibrated angular radius meets a declared tolerance, or abstains when the budget is exhausted. Under exchangeability of complete trajectory bundles, the finite-prefix construction bounds the marginal joint event of executing with excessive unsigned direction error. In Panda simulation with supplied contact and link tracking, using 100 calibration and 100 test geometries, Probe2Act reduces clean interaction cost from 12 to 1.87 queries, with 97.5% versus 98.5% task success. Under combined pose noise, it uses 6.78 queries while both methods achieve 98% success. None of its 800 nominal executions has an initial direction error above 15°. Matched comparisons and ablations identify which estimation and uncertainty choices support these savings. Tracking shifts reduce uncertainty-set coverage despite accurate selected directions, identifying a boundary for transfer to real sensing.

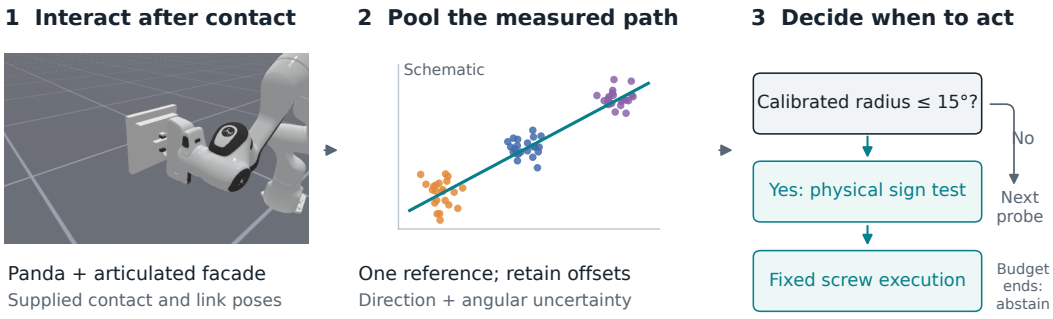

Figure 1: **Spend interactions only until the directional criterion is met.** After contact, the robot probes an unfamiliar articulation, retains the motion across probes in a shared path, and checks a calibrated angular radius. It continues probing while the criterion is unmet, resolves task sign and executes after selection, or abstains at the budget limit. Contact and tracked link poses are supplied; the point-cloud panel is schematic.

1 INTRODUCTION

Opening an unfamiliar drawer or cabinet begins with a physical uncertainty: how can the contacted part move? A handle may belong to a sliding part or a hinged one, and appearance need not reveal the hidden mechanism. Interaction can supply the missing motion evidence, as demonstrated by active articulation estimation and online multimodal systems (Hausman et al., 2015; Buchanan et al., 2026). The practical question is not only how to estimate that motion, but *how many interactions to spend before committing to it*. Exhausting a fixed probing budget can waste interactions when the early measurements are already informative, so stopping too soon can select a poorly supported direction.

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.

Existing methods predict articulation from visual geometry, choose informative actions, and update motion models during manipulation (Eisner et al., 2022; Hausman et al., 2015; Buchanan et al., 2026). Task-oriented planning also recognizes that identification should serve the manipulation objective (Narayanan & Likhachev, 2015). We study a narrower decision within this broader problem: **given a bounded probe sequence and noisy link-motion observations, when does the accumulated evidence support a local direction within a declared tolerance?** The probe sequence is fixed and stopping alone is adaptive. This isolates the cost of deciding to act from the separate problem of choosing the next probing action.

Our starting observation is that weak individual responses need not imply an uninformative history. A probe may move the part briefly and then hold it nearly stationary. Several such responses can reveal a common motion through the displacement *between* them, even when each response contains little variation. Separately centering the responses removes those offsets. Probe2Act instead fits one shared motion path, retaining both within-probe variation and between-probe displacement. It then attaches a working uncertainty scale to the estimated direction and calibrates that scale across every prefix of the acquisition sequence (Figure 1). The robot stops at the first eligible prefix whose radius meets the tolerance, or abstains if no prefix qualifies.

The method uses geometric estimation and offline calibration, without training an action policy. Its intended use is a post-contact identification module: an upstream system has established contact, and the module decides when the observed motion supports a local command. We evaluate this use case in controlled Panda simulation with supplied contact and link tracking. Relative to the same estimator run for all 12 probes, Probe2Act uses 1.87 queries in clean sensing and 6.78 under combined pose noise, with a one-percentage-point observed success deficit and equal success, respectively. Together, the estimator and gate provide an interaction-efficient identification interface whose behavior can be measured separately from the downstream controller.

Our contributions.

(1) Preserve useful motion evidence. A common-path estimator retains displacement across probes and directly estimates an unsigned local screw direction, allowing individually weak responses to contribute jointly.

(2) Make stopping explicit. A finite-prefix calibrated gate converts the estimate into a select, continue, or abstain decision, with a marginal direction-error statement under declared trajectory-exchangeability assumptions.

(3) Measure the practical tradeoff. Matched estimator and stopping comparisons quantify consumed interactions, direction errors, and task success separately. Ablations identify the useful components, competitive simpler alternatives, and limits under tracking shift.

2 Related Work

Predicting and identifying articulated motion. Interactive perception estimates object kinematics through manipulation (Katz & Brock, 2008; Martín-Martín & Brock, 2014; Bohg et al., 2017). FlowBot3D predicts dense articulation flow from point clouds to guide an analytic motion planner (Eisner et al., 2022). Other systems combine initial predictions with observations acquired during execution: Buchanan et al. fuse learned articulation-flow priors and covariance, proprioception, and reaction forces in an online screw-theoretic factor graph (Buchanan et al., 2026). Probe2Act instead conditions on post-contact motion measurements and studies the decision to stop acquiring them. Our matched geometric controls compare estimators under the same sensing and budget, not complete perception-and-control systems.

Choosing interactions and deciding when to stop. Active articulation estimation already selects actions using information-based objectives (Hausman et al., 2015; Eppner et al., 2018), and task-oriented planning identifies mechanisms while completing a goal (Narayanan & Likhachev, 2015). Act2See uses a vision-language planner to choose probe, execute or skip actions while constructing an articulated representation from RGB-D and end-effector observations (Liu et al., 2026). These methods establish uncertainty-aware interaction while we focus on a continuous-direction stopping

criterion within a declared acquisition bank. Active action selection is complementary, but a changed acquisition procedure would require its own calibration justification.

Calibrated decisions and evaluation resources. Split conformal prediction supplies marginal coverage under exchangeability (Angelopoulos & Bates, 2021); conformal risk control extends calibration to general monotone losses (Angelopoulos et al., 2024). Physics-informed and contact-aware scores have also been studied (Gopakumar et al., 2025; Marques et al., 2026). Explore until Confident uses multi-step conformal prediction to stop embodied exploration for question answering (Ren et al., 2024). We apply established finite-prefix reasoning to an unsigned physical direction and evaluate whether the resulting sets permit useful early stopping. SAPIEN/PartNet-Mobility and ManiSkill provide articulated assets and manipulation tasks (Xiang et al., 2020; Gu et al., 2023); newer suites include RoboCasa365 for generalist household manipulation and ArtiBench for vision-language articulated manipulation (Nasiriany et al., 2026; Wu et al., 2025). Our controlled protocol complements these broader evaluations by isolating acquisition cost and directional reliability under matched observations. Appendix F.4 compares method scope; benchmark-scale transfer remains a limitation (Sec. 7).

3 Problem: When Is the Motion Evidence Sufficient?

Given contact and tracked link poses, use $B = 12$ probes (both signs of six Cartesian axes). After k probes, select a unit direction $\hat{\mathbf{s}}_k \in \mathbb{R}^6$, continue or abstain. Approach and grasp acquisition are outside this module.

Motion and notation. Let $\mathbf{T}_G^F \in \mathrm{SE}(3) \subset \mathbb{R}^{4\times4}$ map frame G to F. In fixed interaction frame C, model the link pose as $\mathbf{T}(q) = \exp(q\boldsymbol{\xi}^\wedge)\mathbf{T}_0$, with scalar configuration $q \in \mathbb{R}$, $\mathbf{T}_0 = \mathbf{T}(0)$, and constant screw $\boldsymbol{\xi} = [\mathbf{v}^\top, \boldsymbol{\omega}^\top]^\top \in \mathbb{R}^6\backslash\{\mathbf{0}\}$; $\mathbf{v}, \boldsymbol{\omega} \in \mathbb{R}^3$ are its linear/angular parts. Inverse maps $\wedge, \vee$ convert six-vectors to 4×4 Lie-algebra matrices and back; $\exp, \mathrm{Log}$ are the matrix exponential/local logarithm. Define

$$\mathbf{D}_\ell = \mathrm{diag}(\mathbf{I}_3, \ell\mathbf{I}_3) \in \mathbb{R}^{6\times6}, \qquad \mathbf{s}^\star = \frac{\mathbf{D}_\ell\boldsymbol{\xi}}{\|\mathbf{D}_\ell\boldsymbol{\xi}\|} \in \mathbb{R}^6, \qquad \ell = 0.15 \text{ m}.$$

Here $\mathbf{I}_d$ is the $d \times d$ identity, $\|\cdot\|$ the Euclidean norm, and $\mathbf{s}^\star$ the true unit direction. The inherited scaling $\mathbf{D}_\ell$ weights one radian as 0.15 m; its sensitivity is untested. With task sign unresolved, projective space $\mathbb{RP}^5$ identifies unit vectors $\pm\mathbf{s}$ as one motion line. For unit representatives $\mathbf{s}, \mathbf{t} \in \mathbb{R}^6$, $d(\mathbf{s}, \mathbf{t}) = \arccos(|\mathbf{s}^\top\mathbf{t}|)$ ignores either sign.

Decision target. Selecting at index τ uses $\hat{\mathbf{s}}_\tau$, the estimate after $k = \tau$ probes. Let EXECUTE denote execution after sign resolution. We control

$$\{\text{EXECUTE and } d(\hat{\mathbf{s}}_\tau, \mathbf{s}^\star) > \epsilon_{\mathrm{act}}\}, \qquad \epsilon_{\mathrm{act}} = 15^\circ. \tag{1}$$

The error is unchanged for $-\mathbf{s}^\star$: the tolerance concerns an unsigned line, not sign, success or force. The bound is joint, not conditional on execution. Angles use radians unless marked degrees.

4 Method: Common-Path Evidence and Calibrated Stopping

Probe2Act estimates a direction and working angular scale from pose histories, calibrates the scale over all prefixes, then resolves sign and executes. The estimator, scale construction and acquisition bank are frozen before calibration and testing.

4.1 Preserve displacement between probes

Let $\mathbf{T}_{ij}, \mathbf{T}_r \in \mathrm{SE}(3) \subset \mathbb{R}^{4\times4}$ be pose sample j of probe i and the first probe's middle pose, respectively, in frame C. Form

$$\mathbf{z}_{ij} = \mathbf{D}_\ell\big[\mathrm{Log}(\mathbf{T}_{ij}\mathbf{T}_r^{-1})\big]^\vee \in \mathbb{R}^6. \tag{2}$$

The six-vector $\mathbf{z}_{ij}$ encodes spatial displacement from $\mathbf{T}_r$, retaining between-probe offsets; $\vee$ extracts coordinates, then $\mathbf{D}_\ell$ scales rotation into meters. For noiseless scalar configurations

$q_{ij}, q_r \in \mathbb{R}$, $\mathbf{z}_{ij} = (q_{ij} - q_r)\mathbf{D}_\ell \boldsymbol{\xi} \in \mathbb{R}^6$ if $|q_{ij} - q_r| \, \|\boldsymbol{\omega}\| < \pi$ (no rotation wrapping in the principal logarithm). Samples are therefore collinear. With retained counts n_i, per-probe means $\bar{\mathbf{z}}_i \in \mathbb{R}^6$ and pooled mean $\bar{\mathbf{z}} \in \mathbb{R}^6$, the decomposition

$$\underbrace{\sum_{i=1}^{k}\sum_{j}(\mathbf{z}_{ij} - \bar{\mathbf{z}})(\mathbf{z}_{ij} - \bar{\mathbf{z}})^\top}_{\text{total scatter}} = \underbrace{\sum_{i=1}^{k}\sum_{j}(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)^\top}_{\text{within probes}} + \underbrace{\sum_{i=1}^{k} n_i(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})^\top}_{\text{between probes}} \tag{3}$$

shows the evidence lost by centering each probe separately. Nearly stationary responses can still identify a direction through their differing means. Reference noise, changing screws and logarithm-branch crossings can violate the exact line model.

4.2 Estimate direction and a working angular scale

Remove consecutive duplicate poses and relabel retained $\mathbf{z}_{ij}$ as $\mathbf{z}_a$: $a = 1, \ldots, N_k$ indexes pooled samples, $N_k = \sum_i n_i$. Let $\tilde{\boldsymbol{\delta}}_u \in \mathbb{R}^6$, $u = 1, \ldots, L_k$, be centered/clipped within-probe differences $\mathbf{z}_{i,j+1} - 2\mathbf{z}_{ij} + \mathbf{z}_{i,j-1}$. Starting at scalar weights $w_a = 1$, repeat four times: fit a line, then downweight orthogonal outliers. With mean $\bar{\mathbf{z}}_w = \sum_a w_a \mathbf{z}_a / \sum_a w_a \in \mathbb{R}^6$, the working noise covariance and corrected scatter are

$$\mathbf{Q}_k = \mathrm{FloorPSD}_{10^{-14}}\left(\frac{\sum_{u=1}^{L_k} \tilde{\boldsymbol{\delta}}_u \tilde{\boldsymbol{\delta}}_u^\top}{6\max\{L_k - 1, 1\}}\right) \in \mathbb{R}^{6\times 6}, \tag{4}$$

$$\mathbf{M}_k = \frac{\sum_a w_a(\mathbf{z}_a - \bar{\mathbf{z}}_w)(\mathbf{z}_a - \bar{\mathbf{z}}_w)^\top}{\max\{\sum_a w_a - 1, 1\}} - \mathbf{Q}_k \in \mathbb{R}^{6\times 6}. \tag{5}$$

$\mathrm{FloorPSD}_c$ floors eigenvalues at c. The divisor $6 = 1^2 + (-2)^2 + 1^2$ assumes independent errors and negligible physical second differences. The final $\hat{\mathbf{s}}_k$ is $\mathbf{M}_k$'s leading unit eigenvector; its largest eigenvalue $\lambda_{\max}(\mathbf{M}_k)$ is scalar. Set $\lambda_{+,k} = \max\{\lambda_{\max}(\mathbf{M}_k), 0\}$. Using final weights, we define effective count $n_{\mathrm{eff},k} = \max\{(\sum_a w_a)^2 / \sum_a w_a^2, 2\}$. Require $\sqrt{\lambda_{+,k}} > 10^{-4}$ m and $\lambda_{+,k} / \max\{\mathrm{tr}(\mathbf{Q}_k)/\sqrt{n_{\mathrm{eff},k}}, 10^{-14}\} \geq 3$ (tr: trace). These excitation cutoffs are frozen development heuristics, not significance tests; 10^{-14} guards division (Appendix A).

To estimate directional variability, hold the fitted along-line coordinates fixed and resample orthogonal residuals in circular blocks of four within each probe. Let $\hat{\mathbf{s}}_k^{(m)}$ denote the refitted direction for draw m. The empirical 68th-percentile scale is

$$h_k = \max\left\{1^\circ,\ \mathrm{Quantile}_{0.68}\left(\{d(\hat{\mathbf{s}}_k^{(m)}, \hat{\mathbf{s}}_k)\}_{m=1}^{128}\right)\right\}. \tag{6}$$

This fixed-seed bootstrap preserves the motion profile. Its percentile measures instability, not coverage.

4.3 Calibrate and stop at the first sufficient prefix

For calibration geometry g, $b \in \mathcal{B}$ indexes two mechanisms $\times$ six sensing conditions. With angular scale h_{gbk} from Eq. 6, let $a_{gbk} = d(\hat{\mathbf{s}}_{gbk}, \mathbf{s}^\star_{gb})/h_{gbk}$ for available prefixes, zero otherwise. For N_{cal} bundles and miscoverage $\alpha \in (0, 1)$,

$$A_g = \max_{b\in\mathcal{B},\, 1\leq k\leq B} a_{gbk}, \qquad q_\alpha = A_{(\lceil (N_{\mathrm{cal}}+1)(1-\alpha)\rceil)}. \tag{7}$$

$A_{(r)}$ denotes the r-th smallest score; $A_{(N_{\mathrm{cal}}+1)} = +\infty$. With $N_{\mathrm{cal}} = 100$ and $\alpha = 0.05$, $q_\alpha = A_{(96)} = 3.3259$ is the computed 96th calibration score, not an ablation choice.

For an available test estimate, construct the angular cap

$$\mathcal{C}_k = \{\mathbf{s} \in \mathbb{RP}^5 : d(\mathbf{s}, \hat{\mathbf{s}}_k) \leq r_k\}, \qquad r_k = \min\{\pi/2, q_\alpha h_k\}. \tag{8}$$

An unavailable estimate receives $\mathcal{C}_k = \mathbb{RP}^5$: every direction remains possible, so it cannot be selected. The first available prefix k satisfying $r_k \leq \epsilon_{\mathrm{act}}$ defines τ and selects $\hat{\mathbf{s}}_\tau$. Otherwise probe again, or abstain at B. Sign resolution may veto execution. No cap intersection or extra verification is used.

Finite-prefix guarantee. Under exchangeable complete bundles, frozen estimation/acquisition/measurement procedures and constant-screw targets, split-conformal ranking gives $\Pr(A_{\text{test}} \le q_\alpha) \ge 1-\alpha$, ensuring simultaneous prefix coverage. Let τ_b be combination b's first eligible prefix (infinity if none), and EXECUTE_b its execution event. Then

$$\Pr(\exists b \in \mathcal{B},\, k \le B : \text{EXECUTE}_b,\ \tau_b = k,\ d(\hat{\mathbf{s}}_{bk}, \mathbf{s}_b^\star) > \epsilon_{\text{act}}) \le \alpha. \tag{9}$$

Condition $\tau_b = k$ checks only the selected prefix. Observation-dependent stopping needs no prefix independence. The bound is marginal over calibration/test bundles, not conditional on either calibration or execution; sign, success, sensing shifts and changed probing policies are excluded. Appendix A gives the proof; Appendix B.4 audits acquisition.

4.4 Resolve task sign and execute

At the known closed boundary, select a sign $\sigma \in \{-1, +1\}$ from five-frame averaged excursions along $\hat{\mathbf{s}}_\tau$ relative to an averaged noisy closed reference. Retain earlier excursions when later probes return to the boundary. If none exceeds a noise-scaled threshold, test up to two opposite physical trial steps; unresolved sign causes abstention. This shared sign procedure is outside Eq. 9.

The fixed executor accumulates 0.01-m commensurate spatial increments along $\sigma\hat{\mathbf{s}}_\tau$, without feedback pacing or re-identification. Reference acquisition, probes, sign trials and execution share 400 control steps. Appendix A specifies the sign threshold, frame composition and termination.

5 Evaluation Setup

We evaluate 100 calibration and 100 test geometries in SAPIEN/ManiSkill (Xiang et al., 2020; Gu et al., 2023), each pairing sliding and hinged mechanisms with identical passive RGB. The frozen method uses supplied contact, link poses, an interaction frame and progress-based termination. All methods share a 12-probe bank and a 400-control-step episode budget. Calibration covers clean sensing, 1/5-mm translation noise, $0.5/2^\circ$ rotation noise and combined 1-mm/0.5° noise. Full simulator executions cover clean and moderate noise; severe noise and tracking shifts use identification replay.

The controls isolate stopping (common-path fixed-12), estimation (rotation-aware, noise-normalized geometric-v2 fixed-12) and calibration (the $2h_k$ heuristic). Executions use independent simulator initialization; geometric-v2 and replay ablations are post-hoc comparisons on the same test geometries. We report consumed queries, errors above 15°, abstention, coverage and success (final normalized progress ≥ 0.8). Execution queries include probing and sign trials; replay counts identification probes, including budget-exhausted abstentions. Descriptive paired 95% intervals resample geometry clusters, and coverage intervals use geometry bundles. Appendix B gives the full protocol; Appendix B.4 records cohort provenance and initialization checks.

6 Experiments

We evaluate interaction savings, matched-budget direction accuracy, calibration, alternative scores and component ablations, followed by tracking-shift diagnostics. The primary method remains frozen.

6.1 How many interactions does stopping save?

Probe2Act consumes fewer queries without exhausting the common bank. It uses 1.87 queries in clean sensing and 5.93–6.78 under moderate noise, compared with 12 for both fixed-budget estimators (Table 1). Against common-path fixed-12, the reductions are 84.4% clean and 43.5–50.6% under moderate noise. Task success remains 95.0–98.0%, but these savings do not imply superior completion: clean success is one percentage point lower and combined-noise success is equal. The paired combined-noise difference is zero (95% interval -2.0 to $+2.0$ points); under translation noise it is -3.0 points (-6.5 to $+0.5$). These are observed tradeoffs, not equivalence or noninferiority claims (Appendix C.3).

Table 1: **Fewer queries with separate accuracy and success measures.** Each condition contains 200 simulator episodes from 100 test geometries under the corrected protocol. Queries include probing and sign trials; violations are initial unsigned errors above $15°$.

Condition	Method	Success/200	Violations	Mean queries
Clean	**Probe2Act**	195	0	1.87
	Common-path fixed-12	197	0	12.00
	Geometric-v2 fixed-12	197	0	12.00
	Uncalibrated $2h_k$	195	0	1.87
Translation 1 mm	**Probe2Act**	190	0	5.93
	Common-path fixed-12	196	0	12.00
	Geometric-v2 fixed-12	197	41	12.00
	Uncalibrated $2h_k$	194	3	3.53
Rotation $0.5°$	**Probe2Act**	196	0	6.28
	Common-path fixed-12	198	0	12.00
	Geometric-v2 fixed-12	197	19	12.00
	Uncalibrated $2h_k$	195	5	4.01
Combined	**Probe2Act**	196	0	6.78
	Common-path fixed-12	196	0	12.00
	Geometric-v2 fixed-12	197	23	12.00
	Uncalibrated $2h_k$	195	0	5.21

All 800 corrected Probe2Act runs select initial directions within $15°$, with no sign abstentions. Remaining failures occur during execution. Figure 2 shows successful sliding and hinged motions alongside a failure despite an accurate initial direction. Figure 3 displays the matched cost/success operating points.

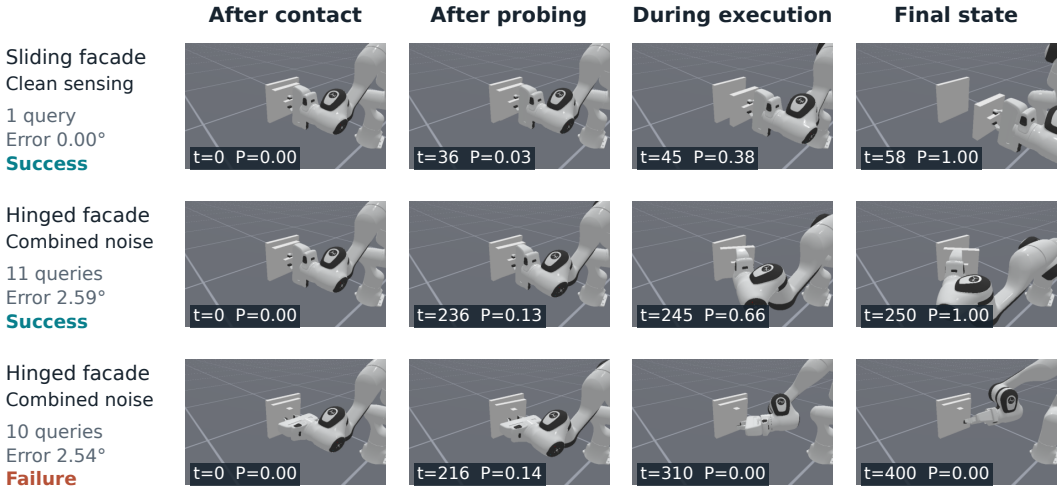

Figure 2: **One stopping rule, different interaction costs.** Clean sliding, noisy hinged success and noisy hinged failure after accurate identification. Captures reproduce existing episode records; t counts control steps and P is joint progress. Selection and provenance are in Appendix F.2.

6.2 Are the directions accurate at a matched budget?

Common-path estimation explains the fixed-budget direction advantage. Both fixed-12 estimators receive the same observations. Common-path records no direction violations; geometric-v2 records 41, 19 and 23 per 200 episodes under translation, rotation and combined noise, all on revolute mechanisms, despite 197/200 task successes in each condition. This comparison isolates estimation from adaptive stopping.

Matched-prefix replay uses only the first k probes. Prismatic median errors approach sub-degree levels by $k \approx 2$–3. Under 1-mm translation noise, the revolute common-path estimate has full availability and $6.9°$ median error at $k = 3$; after intermediate availability fluctuations, it reaches full

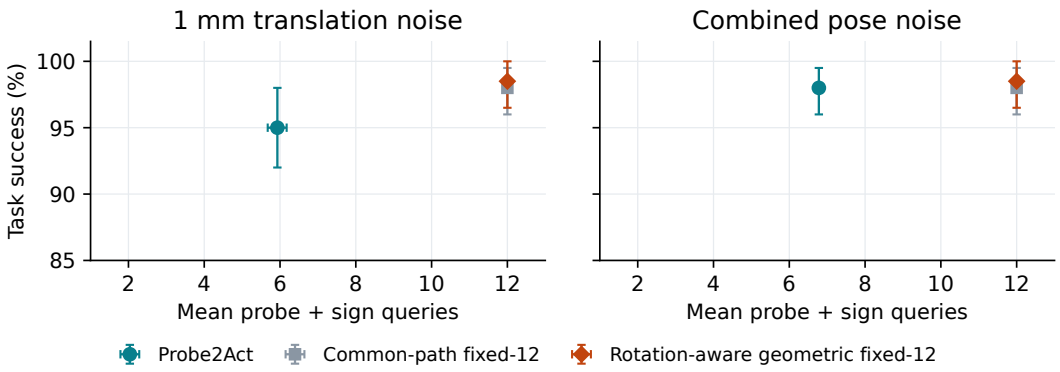

Figure 3: **Fewer interactions at high observed task success.** Matched success versus probing/sign queries.

availability and 1.0° error at $k = 12$. Geometric-v2 is available throughout but has larger intermediate errors (Table 2). Availability, accuracy and calibrated stopping eligibility remain distinct. Appendix C.2 gives all mechanisms and conditions. Appendix C documents both geometric fits;

Table 2: Selected matched-prefix curve, revolute mechanisms, 1 mm translation noise (100 trajectories). Cells show fraction available (median angular error in degrees, computed only over the available trajectories at that prefix; n/a if none are available). Table S1 (Appendix C.2) gives both mechanisms and all four primary conditions at $k \in \{1, 6, 12\}$; the underlying 19,200-row file gives every prefix.

Estimator	$k = 1$	$k = 3$	$k = 6$	$k = 9$	$k = 12$
Common-path	0.01 (82.6°)	1.00 (6.9°)	0.97 (5.5°)	0.96 (4.4°)	1.00 (1.0°)
Geometric-v2	1.00 (76.3°)	1.00 (82.1°)	1.00 (82.4°)	1.00 (81.7°)	1.00 (4.3°)

these are matched pose-based controls, not full learned or multimodal systems.

6.3 What does calibration change?

Calibration selects a direction-error/cost tradeoff, not an improvement in every outcome. The $2h_k$ heuristic uses the same estimate and working scale but a smaller multiplier. Under translation and rotation noise, it spends fewer queries and records three and five direction violations, respectively, whereas Probe2Act records none (Table 1). Their clean outcomes coincide, and the heuristic completes more translation-noise tasks. Only one such heuristic multiplier was confirmed in simulator execution; the next section compares additional calibrated scores and levels in replay.

Full-prefix coverage is 92/100 nominal geometry bundles at requested 95% coverage (Wilson interval 85.0–95.9%). This interval includes the target but does not establish precise calibration. Simultaneous coverage is stricter than direction accuracy at the selected prefix, so coverage misses can coexist with zero selected-direction violations. Severe-noise replay often abstains and does not supply physical task-success evidence.

6.4 Is the bootstrap better than simpler calibrated scores?

We keep the direction estimator fixed and independently calibrate three alternative scales: $1\,\mathrm{rad}/\sqrt{k}$, a noise-to-excitation proxy using $\mathbf{Q}_k$, $\lambda_{+,k}$ and $n_{\mathrm{eff},k}$, and the change from the current estimate to the last prior available estimate. A fourth control applies the stability scale to geometric-v2. Each score uses the same calibration bundles; comparisons concern replay selection and cost, not task execution. Appendix D defines the scales, unavailable-prefix conventions and three calibration levels. **The bootstrap has condition-dependent value, not uniform superiority.** At $\alpha = 0.05$, the excitation proxy uses fewer probes under 1-mm translation noise (5.07 versus 5.93), with equal eligibility and no observed violations. The bootstrap selects 131/200 directions under severe rotation, compared with 100/200 for the proxy, again with no violations. Direction stability incurs several nominal direction errors. Thus a cheaper scale can be competitive, but it does not match every operating point (Table 3).

Table 3: **Calibrated scores trade eligibility against interaction cost.** Independent calibration on 100 geometry bundles; evaluation on the same 100 test geometries at $\alpha = 0.05$. E: eligible directions/200; V: errors above 15°; K: mean identification probes, including budget-exhausted abstentions. Replay excludes execution and sign trials. Probe-count and geometric-stability eligibility is zero at this level (Appendix D).

	Bootstrap (frozen)			Excitation proxy			Direction stability		
Condition	E	V	K	E	V	K	E	V	K
Clean	200	0	1.83	200	0	1.83	200	0	2.83
1 mm translation	200	0	5.93	200	0	5.07	200	2	5.68
5 mm translation	100	0	10.02	102	0	11.15	125	2	9.70
0.5° rotation	200	0	6.28	200	0	6.46	200	1	6.87
2° rotation	131	0	11.11	100	0	11.28	120	0	9.78
Combined	200	0	6.78	200	0	6.79	199	1	7.61
Nominal bundle coverage		92/100			92/100			93/100	
Separate tracking-shift diagnostic (outside the nominal guarantee)									
Correlated-shift V/200		0/200			133/200			5/200	

Under the correlated-noise shift, outside the guarantee, bootstrap and excitation proxy are both eligible on all 200 trajectories. At $\alpha = 0.05$, the proxy stops after 2.44 mean probes and records 133 direction errors; the bootstrap uses 6.825 probes and records none. The proxy's error proportion ranges from 51% to 76.5% across the three tested levels. This is a measured shift-specific advantage, not a new coverage guarantee or evidence that temporal blocking specifically causes it.

The probe-count schedule never qualifies at $\alpha \in \{0.025, 0.05\}$; at $\alpha = 0.10$, it selects 1,799/1,800 directions with one error, always at $k = 12$. Geometric stability remains ineligible because its calibrated multiplier times the 1° floor exceeds the stopping tolerance. Appendix D details these score–floor–calibration effects.

Single-threaded computation is 39.1 ms for the core fit and 48.1 ms with bootstrap, not end-to-end task latency. Appendix D gives timing and full replay summaries.

6.5 Which design choices support early stopping?

Table 4: Post-hoc replay with separate recalibration. Cells show eligible trajectories/200 and mean identification probes, including abstentions but excluding sign trials. Coverage uses 100 six-case geometry bundles. With zero selected-direction errors across 1,800 decisions per variant, retaining between-probe displacement and the noise model chiefly improves eligibility and cost.

Variant	q_α	Combined	2° rotation	5 mm translation	Bundle cov.
Full method	3.326	200 (6.78)	131 (11.11)	100 (10.02)	92/100
Per-probe centering	4.491	105 (11.27)	0 (12.00)	13 (11.95)	94/100
Uniform weights	3.326	200 (6.78)	131 (11.11)	100 (10.02)	92/100
$\mathbf{Q} = \mathbf{0}$ throughout	90.0	0 (12.00)	0 (12.00)	0 (12.00)	97/100
Pooled i.i.d. resampling	4.560	200 (6.79)	71 (11.59)	100 (11.27)	98/100
All four changes	6.913	42 (11.91)	0 (12.00)	0 (12.00)	96/100

Table 5: Post-hoc replay coverage decomposition: 200 trajectories nested within 100 geometries per condition, all with an eligible direction. Full-bank coverage checks all 12 prefixes; selected-cap coverage checks the stopping prefix. Some selected caps miss the truth despite center errors below 15°: accurate directions do not establish calibrated sets.

Condition	Full-bank coverage	Selected-cap coverage	Error $> 15^\circ$
Correlated noise	93/200	196/200	0/200
Initial outlier	1/200	200/200	0/200
Dropout	200/200	200/200	0/200

Preserve motion between probes. Removing the between-probe displacement term reduces combined-noise eligibility from 200/200 to 105/200 and severe-rotation eligibility from 131/200

to zero after separate recalibration (Table 4). The same acquired trajectories support fewer decisions when their offsets are discarded.

Retain the fitted-scatter noise correction. Removing all uses of $\mathbf{Q}$ makes every calibrated radius too wide for selection. Removing only its fitted-scatter correction also yields zero nominal eligibility ($q_\alpha = 65.14$; Table S5). The scale reaches its 1° floor on 99.75% of valid prefixes, versus 20.2% for the default, and geometry-level maximum scores become consistently large. Paired-valid-prefix errors change more modestly, but do not isolate error or availability changes at the score-determining prefixes. The result concerns this estimator–gate combination.

Preserve probe-specific residual structure. Pooled independent resampling reduces severe-rotation eligibility to 71/200. Independent resampling *within* each probe instead yields 141/200 versus the default's 131/200, but introduces one nominal 5-mm-translation direction error (Table S5). Probe-specific residual structure has the clearer demonstrated benefit; temporal blocks show little additional benefit in this comparison. Uniform weighting leaves nominal decisions unchanged. The six original ablations record no selected-direction violations; additional sensitivity tests introduce one with block length two and three with doubled $\mathbf{Q}$, the latter under first-frame-outlier shift. Appendix E gives the definitions, isolated controls and parameter sweep. No default was retuned, and these replay ablations do not establish task-success effects.

6.6 Where does calibration stop transferring?

Direction accuracy and set coverage differ under tracking shift. Correlated noise gives 93/200 full-bank and 196/200 selected-cap coverage, despite no selected-direction violations. The initial outlier gives 1/200 full-bank and 200/200 selected-cap coverage (Table 5). These shifts are outside Eq. 9's assumptions. Appendix F.1 gives the decomposition; historical physical-shift runs remain separate.

7 Conclusion

In summary, Probe2Act yields two fundamental insights. First, probing effort can follow accumulated evidence: preserving displacement between probes makes weak responses jointly informative, while calibration links stopping to a declared directional tolerance. This combination saves interactions in our controlled setting. Second, directional accuracy, uncertainty-set coverage, and task completion require separate evaluation. Tracking shifts can preserve useful estimates while undermining coverage, and accurate identification can still be followed by execution failure.

Discussion and future work. The controlled post-contact study uses a fixed probe bank, one-DoF mechanisms and supplied contact and link tracking. Broader mechanisms and real visual sensing are the next validation targets, using the same cost–error–coverage protocol within existing manipulation suites (Wu et al., 2025; Nasiriany et al., 2026). Future integration could constrain post-contact VLA commands using the identified direction (Kim et al., 2024), with compatible sensing and calibration. The statistical guarantee retains the assumptions in Sec. 4.3; Appendix F.4 collects the detailed scope and deployment considerations.

AI use statement

In this work, we used generative AI tools (i.e. ChatGPT) for revising passages, and correcting grammar. All research ideas, methods, experiments, analyses, and conclusions are the authors' own. All AI-assisted text was reviewed by the authors. The authors take full responsibility for all content of this paper.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S. Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017. doi: 10.1109/TRO.2017.2721939.

Russell Buchanan, Adrian Röfer, João Moura, Abhinav Valada, and Sethu Vijayakumar. Online estimation and manipulation of articulated objects. *Autonomous Robots*, 50(1):13, 2026. doi: 10.1007/s10514-025-10233-4.

Ben Eisner, Harry Zhang, and David Held. FlowBot3D: Learning 3D articulation flow to manipulate articulated objects. In *Robotics: Science and Systems*, 2022. URL https://arxiv.org/abs/2205.04382.

Clemens Eppner, Roberto Martín-Martín, and Oliver Brock. Physics-based selection of informative actions for interactive perception. In *IEEE International Conference on Robotics and Automation*, pp. 7427–7432, 2018. doi: 10.1109/ICRA.2018.8460596.

Vignesh Gopakumar, Ander Gray, Lorenzo Zanisi, Timothy Nunn, Daniel Giles, Matt Kusner, Stanislas Pamela, and Marc Peter Deisenroth. Calibrated physics-informed uncertainty quantification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20103–20141. PMLR, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

Karol Hausman, Scott Niekum, Sarah Osentoski, and Gaurav S. Sukhatme. Active articulation model estimation through interactive perception. In *IEEE International Conference on Robotics and Automation*, 2015. doi: 10.1109/ICRA.2015.7139655. URL https://karolhausman.github.io/pdf/icra15articulation.pdf.

Dov Katz and Oliver Brock. Manipulating articulated objects with interactive perception. In *IEEE International Conference on Robotics and Automation*, pp. 272–277, 2008. doi: 10.1109/ROBOT.2008.4543220.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Yifeng Liu, Yuzhen Chen, Bingyang Wang, Chengqi Li, Jingyi Lu, Han Li, Kaichen Zhou, Mengyu Wang, and Fangneng Zhan. Act to see: Structured articulated representations through robot interaction. In *Robotics: Science and Systems Workshop on Foundation Models for Robot Planning (FM4RoboPlan)*, 2026. URL https://openreview.net/forum?id=sghH6peEK8.

Luís Marques, Kristian Popov, and Dmitry Berenson. Particle-based conformal prediction for contact-aware uncertainty calibration in stratified configuration spaces, 2026.

Roberto Martín-Martín and Oliver Brock. Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2494–2501, 2014. doi: 10.1109/IROS.2014.6942902.

Venkatraman Narayanan and Maxim Likhachev. Task-oriented planning for manipulating articulated mechanisms under model uncertainty. In *IEEE International Conference on Robotics and Automation*, pp. 3095–3101, 2015. doi: 10.1109/ICRA.2015.7139624. URL https://publications.ri.cmu.edu/task-oriented-planning-for-manipulating-articulated-mechanisms-under-model-uncertainty.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.04356.

Allen Z. Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024. URL https://arxiv.org/abs/2403.15941.

Yuhan Wu, Tiantian Wei, Shuo Wang, ZhiChao Wang, Yanyong Zhang, Daniel Cremers, and Yan Xia. ArtiBench and ArtiBrain: Benchmarking generalizable vision-language articulated object manipulation. *arXiv preprint arXiv:2511.20330*, 2025. URL https://arxiv.org/abs/2511.20330.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.