Uncertainty-Aware Execution Monitoring and Human Handover for LLM-Generated Robot Programs

Archit Jain
University of Washington, Seattle

Abstract

As large language models are increasingly used to synthesize executable robot programs from natural-language descriptions, deciding when to interrupt execution and request human help becomes a problem of probabilistic state estimation grounded in calibrated sensor models – one that has received comparatively little attention relative to the program-synthesis side of the literature. We present a calibrated, recursive Bayesian estimator that tracks grasp success during execution of LLM-generated robot programs on a real Franka manipulator. The estimator fuses a binary visual-question-answering (VQA) sensor with a continuous gripper-width residual, both calibrated from labeled trials via maximum-likelihood and least-squares system identification. Filter consistency is monitored via a chi-squared test on the gripper-width innovation, adapted from the Normalized Innovation Squared (NIS) consistency check used in linear Kalman filtering. On a held-out set of 19 tabletop pick-and-place trials with parameters frozen from a 26-trial training set, the NIS-based detector and an existing rule-based grasp-check baseline catch complementary subsets of grasp failures – mid-execution slips and clean-empty grasps respectively – and their disjunction detects all 11 grasp failures at zero false positives. Calibration additionally revealed that the gripper-width encoder is effectively deterministic at primitive-event sampling rates and that the GPT-based VQA sensor is largely non-informative in this configuration, both reported as findings rather than corrected for.

1 Introduction

Large language models (LLMs) are increasingly used to generate executable robot programs from natural-language task descriptions [2–4]. The stochastic nature of these models, however, makes runtime uncertainty difficult to quantify, and the perception sensors typical of modern robot stacks – vision-language models for binary task verification, learned grasp samplers, and geometric primitives – add noise that is opaque and often miscalibrated. For a manipulation program to fail gracefully and hand control back to a human at the right moment, a principled probabilistic estimate of *whether the manipulation is currently succeeding* is needed, derived from sensors whose noise has been characterized rather than assumed.

This work develops such an estimator for the specific case of tabletop grasping. The state is a scalar belief that the target object is currently held in the gripper; the sensors are a GPT-based VQA call and the gripper-width encoder; the estimator is a recursive Bayesian filter that updates at every primitive-completion event in the generated program; and the failure trigger is a chi-squared consistency test on the gripper-width residual, adapted from the standard NIS check [1]. The system runs on a real 7-DoF Franka arm with a Robotiq 2F-85 gripper, instrumented through an existing CLI that synthesizes Python from natural-language tasks and dispatches ROS primitives.

An initial design for this work considered a 20-dimensional Kalman filter over joint, end-effector, and object pose evaluated in simulation, together with a pre-execution LLM-uncertainty layer based on abstract syntax tree distance over sampled programs. Two findings during integration on the actual hardware motivated a refined scope. First, calibration of the available sensors revealed that the gripper-width encoder is effectively deterministic at the timescale of primitive-completion events

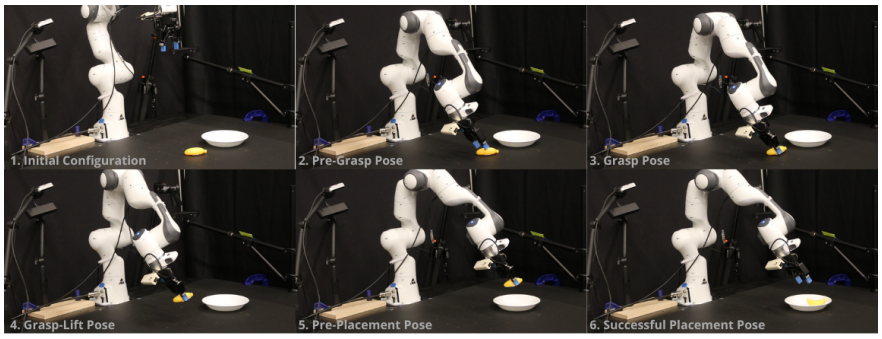


Figure 1: Six-stage manipulation sequence on the experimental setup used for all trials: a banana on the right of the workspace, a bowl on the left, single-object configuration. From left to right, top to bottom: initial configuration, pre-grasp pose, grasp pose, post-lift pose, pre-placement pose, and successful placement pose. All 45 trials reported below use this same physical setup.

($\hat{R} = 0$ from least squares), and that the VQA sensor returns "grasped" nearly regardless of ground truth ($\hat{\text{FPR}} = 1.0$ on missed-grasp trials) – both findings that significantly reshape what kind of estimator is well-posed on this hardware. Second, the dominant runtime failures observed on the real platform are physical (slip, missed grasp, mis-placement) rather than program-structural, making LLM program-variance an unproductive target. The estimator presented below is what calibration drove the design toward: a one-dimensional belief, a recursive Bayes update fusing the two characterized sensors, and a NIS test focused on the signal that calibration revealed to be informative.

The headline empirical result is one of *sensor complementarity*: on 19 held-out trials with parameters frozen, the NIS-based width tracker catches all 6 slip failures and zero clean-missed; the existing rule-based grasp-check catches all 5 clean-missed and zero slips; and their disjunction detects 11/11 grasp failures at zero false positives. The remainder of the paper covers the platform and problem formulation (§2), the recursive estimator and NIS test (§3), the system identification of both sensors (§4), and the experimental results on training and held-out data (§5), before concluding with limitations (§6).

2 Background and Problem Formulation

Platform. The execution stack accepts a natural-language task description, prompts an LLM to synthesize a Python program over a fixed set of robot primitives, optionally allows the operator to edit the program, and then executes it on a Franka Emika Panda arm with a Robotiq 2F-85 parallel-jaw gripper. Each primitive is implemented as a ROS service. Control primitives (`set_gripper_width`, `move_ee_to_pose`, `execute_waypoint_trajectory`, `rotate_wrist`, `reset_robot`, and guarded variants thereof) are blocking calls that return after motion completes. Perception primitives include Grounded SAM2-based segmentation [5], AnyGrasp-based grasp sampling [6], AnyPlace placement prediction, DIFT keypoints, and a GPT-based VQA call. Proprioception primitives expose the current end-effector pose, joint state, and gripper aperture in metres. Figure 1 shows the physical setup used in this study: a banana on the right of the workspace and a bowl on the left, viewed from the wrist and overhead cameras.

Sensors available for runtime monitoring. Within the generated program, two sensors are accessed at every primitive boundary by an instrumented logger: `get_current_gripper_width`

(continuous, metres) and `get_vqa_response` (a natural-language answer to a query, parsed to a binary truth value for the question "Is the banana grasped?"). A third signal is the heuristic `verify_grasp` primitive, which already exists in the system: it returns `False` if the gripper is fully closed and otherwise queries a vision-language model with the wrist-camera image. This is treated as the baseline detector in §5.

Failure modes. Across 45 logged trials, four outcome classes were observed: *success*, *slip* (the object is grasped at the close event but lost during lift or transit), *missed* (the gripper closes on empty air, or on the edge of the object such that the grasp never forms), and *wrong placement* (the object is held throughout but ends up outside the target). The detection target in this work is the union of *slip* and *missed* – collectively, "grasp not held throughout execution" – since these are the failures a grasp-belief estimator can in principle address. *Wrong placement* is correctly outside the scope: a held grasp ending in the wrong location is not a grasp failure.

Problem statement. Given a sequence of primitive-completion events $k = 0, 1, \ldots, K$ generated by execution of an LLM-synthesized program, and a stream of sensor observations available at each event, maintain a recursive probabilistic estimate $b_g(k) \in [0, 1]$ of the current grasp status, and trigger a handover to the operator whenever the gripper-width innovation sequence is statistically inconsistent with the calibrated grasp model. The estimator parameters must be calibrated by system identification on a held-out-disjoint training set and then frozen prior to evaluation.

3 Methodology

3.1 State, observation models, and recursive update

The state at event k is the scalar belief

$$b_g(k) \triangleq \Pr[\,\text{object is currently held at event } k\,] \in [0, 1], \tag{1}$$

indexed by primitive-completion events rather than wall-clock time. Because the underlying program executes as a sequence of blocking ROS calls with stationary intervals between them, this event-indexed formulation matches the natural rhythm of the system and avoids the asynchronous-logging machinery that a high-rate time-indexed filter would require.

The process model is trivial: in the absence of intermediate dynamics between primitive events, the prior at k equals the posterior at $k-1$, i.e. $b_g(k \mid k-1) = b_g(k-1)$. All inference happens in the measurement update.

VQA sensor (binary). When the generated program calls the grasp VQA query, the parsed answer $z_v(k) \in \{0, 1\}$ is modelled with parameters $\mathrm{FPR}, \mathrm{FNR} \in [0, 1]$:

$$\Pr[z_v = 1 \mid g = 1] = 1 - \mathrm{FNR}, \quad \Pr[z_v = 1 \mid g = 0] = \mathrm{FPR}, \tag{2}$$

with the corresponding $z_v = 0$ probabilities determined by complementation.

Gripper-width sensor (continuous). Let $w(k)$ be the gripper aperture (metres) recorded at event k, and let w_b be the baseline aperture sampled at the first event immediately following the close primitive of the grasp attempt. The residual $r_w(k) \triangleq w_b - w(k)$ is positive whenever the fingers have closed further than at baseline, the signature of slip. Under the calibrated noise model,

$$r_w(k) \mid g{=}1 \sim \mathcal{N}(0, R), \qquad r_w(k) \mid g{=}0 \sim \mathcal{N}(w_b, R), \tag{3}$$

where the second branch reflects the empirical fact that an empty gripper closes to a width near zero, producing a residual near w_b that is very improbable under $\mathcal{N}(0, R)$ for any reasonable R.

Recursive Bayes update. Working in posterior-odds form for numerical stability,

$$o(k) \triangleq \frac{b_g(k)}{1 - b_g(k)}, \qquad o(k) = o(k-1) \cdot \prod_{i \in \mathcal{O}(k)} \Lambda_i(k), \tag{4}$$

where $\mathcal{O}(k)$ is the set of observations available at event k and the likelihood ratios $\Lambda_i = \Pr[\cdot \mid g{=}1] / \Pr[\cdot \mid g{=}0]$ are

$$\Lambda_v(k) = \begin{cases} (1 - \mathrm{FNR})/\mathrm{FPR}, & z_v(k) = 1 \\ \mathrm{FNR}/(1 - \mathrm{FPR}), & z_v(k) = 0 \end{cases}, \quad \Lambda_w(k) = \frac{\mathcal{N}(r_w(k);\, 0, R)}{\mathcal{N}(r_w(k);\, w_b, R)} = \exp\Big(\tfrac{w_b}{R}\big(r_w(k) - \tfrac{w_b}{2}\big)\Big).$$

The belief is clipped to $[\varepsilon,\, 1 - \varepsilon]$ with $\varepsilon = 10^{-3}$ to keep odds finite.

3.2 NIS consistency test

The NIS check for a linear-Gaussian filter exploits the fact that, under correct dynamics and noise covariance, the innovation squared and Mahalanobis-normalized by the innovation covariance is chi-squared distributed with degrees of freedom equal to the measurement dimension [1]. The same logic applies here: under the hypothesis "currently grasped and width noise model correct," the per-event statistic

$$\mathrm{NIS}(k) = \frac{r_w(k)^2}{R} \tag{5}$$

is distributed as $\chi^2(1)$. The 95% upper quantile $\chi^2_{0.95}(1) \approx 3.841$ defines the per-event trigger threshold; exceeding it at any post-baseline event indicates that the observed width is inconsistent with the held-grasp model and triggers a handover to the operator.

4 System Identification

4.1 VQA error rates via maximum likelihood

On the training set (§5), the grasp-VQA call returned "yes" on 26/26 trials, including all 8 trials in which the ground-truth label indicated a missed grasp. The maximum-likelihood estimates are therefore

$$\widehat{\mathrm{FPR}}_{\mathrm{MLE}} = 8/8 = 1.000, \qquad \widehat{\mathrm{FNR}}_{\mathrm{MLE}} = 0/18 = 0.000,$$

with 95% Wilson score intervals $\mathrm{FPR} \in [0.68, 1.00]$ and $\mathrm{FNR} \in [0.00, 0.18]$. These point estimates are degenerate for the Bayes update in (4): Λ_v collapses to 1 when $z_v = 1$, providing no information, and to 0 when $z_v = 0$ (an outcome that never occurs in the calibration data). To keep the update numerically well-defined, the parameters used by the estimator are Laplace-smoothed with a single pseudocount in each cell:

$$\widehat{\mathrm{FPR}}_{\mathrm{Lap}} = \frac{8+1}{8+2} = 0.900, \qquad \widehat{\mathrm{FNR}}_{\mathrm{Lap}} = \frac{0+1}{18+2} = 0.050.$$

The substantive finding – that VQA is essentially non-informative in this configuration – is preserved by the smoothed values and is reported as a calibration result, not corrected for.

Scenario	Residual (mm)	NIS at $\sigma_w \in \{0.5, 1, 2\}$ mm		
		0.5 mm	1 mm	2 mm
Slip (held → released)	23	2116	529	132
Missed (empty grip from close)	0	0	0	0
Clean success	0	0	0	0
Within-grasp settling tremor	0.5	1.00	0.25	0.06

Table 1: NIS robustness to the noise-floor choice. The slip case sits well above the 3.841 threshold across the entire range; the settling tremor stays well below it. The detection conclusion is insensitive to the precise value of σ_w.

Parameter	Raw estimate	Value used by estimator
FPR	1.000 (MLE; Wilson 95%: [0.68, 1.00])	0.900 (Laplace)
FNR	0.000 (MLE; Wilson 95%: [0.00, 0.18])	0.050 (Laplace)
R (m^2)	0 (least squares on M=12 residuals)	10^{-6} (floor: σ_w=1 mm)

Table 2: Calibrated sensor parameters. Raw estimates are reported as system-identification findings; smoothed/floored values are the frozen parameters used in all downstream evaluation.

4.2 Width residual variance via least squares

From the 6 training trials labelled "success" (object held throughout), let $\mathcal{H}$ denote the set of width residuals $r_w(k)$ at events strictly between the close primitive and the subsequent open primitive. The least-squares variance estimate

$$\widehat{R}_{\mathrm{LS}} \;=\; \frac{1}{|\mathcal{H}| - 1} \sum_{r_w \in \mathcal{H}} r_w^2 \;=\; 0$$

indicates that the encoder-reported width is bit-identical across all held events of all successful trials – the sensor is, to within its quantization, deterministic at primitive-event sampling rates. A purely encoder-quantization noise model ($R \approx q^2/12 \approx 8 \times 10^{-10}\,\mathrm{m}^2$) would render the NIS test pathologically sensitive: a 0.5 mm settling motion of the object inside the gripper would produce NIS ≈ 300. The variance that matters for the estimator is not encoder resolution but the physical phenomenon of *micro-motion within a maintained grasp* – finger-pad compression, sub-millimetre re-seating of the object during transit – which lives on the millimetre scale. A physical floor $\sigma_w = 1\,\mathrm{mm}$ is therefore imposed:

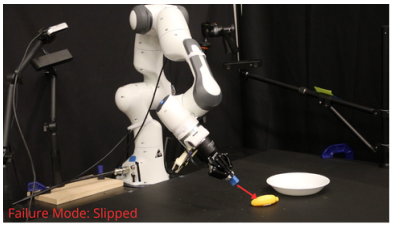
$$R \;=\; \max(\widehat{R}_{\mathrm{LS}},\, \sigma_w^2) \;=\; 10^{-6}\,\mathrm{m}^2.$$

The robustness of the NIS test to the exact value of σ_w is reported in Table 1: a 10× variation in σ_w changes the slip-case NIS by a factor of 100 but keeps the slip case three orders of magnitude above the 3.841 threshold.

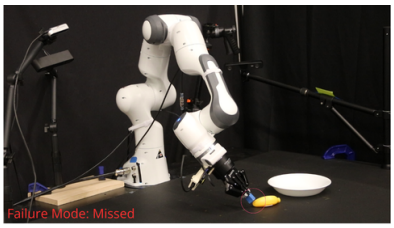
5 Numerical Results

5.1 Experimental design

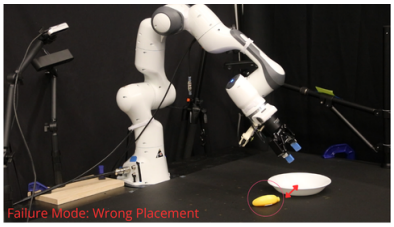
A single task – "Pick up the banana and place it in the bowl" – was used throughout, with the same single-object setup of Figure 1. Trials were collected in two phases: a 26-trial training phase used for



(a) Slip: object held at close, lost mid-lift. Width: $0.032 \rightarrow 0.009$ m.



(b) Missed grasp: gripper closes on empty space after the object is moved. Width: ~ 0.008 m.



(c) Wrong placement: object held throughout, deposited outside the bowl.

Figure 2: Representative frames of the three deliberately induced failure modes. Red markers indicate the location of the banana at the moment of failure; green markers indicate the intended target.

calibration, and a subsequent 19-trial held-out phase collected after the parameters in Table 2 were committed to a frozen JSON artifact (`calibration.json`) and the analysis pipeline was tagged at the corresponding commit. In each trial, the operator either ran the LLM-generated program with no interference (success) or deliberately induced one of three failure modes by physical intervention: (i) *slip*, by manually opening the Robotiq gripper a few millimetres during the lift phase to dislodge the object; (ii) *missed*, by sliding the banana 3–5 cm laterally after the segmentation-based grasp pose had been computed, so that the gripper closes on empty space or on the banana's edge; (iii) *wrong placement*, by translating the bowl by ~ 10 cm after the program began, so that the cached placement pose lands the banana outside the target. Outcomes were labelled at trial-end via an interactive prompt that stored a structured ground-truth tuple in each trial JSON. Figure 2 shows representative frames of each failure mode.

5.2 Calibration replay on the held-out set

Figure 3 summarizes the calibration outputs on the held-out trials, using the frozen parameters from Table 2. The left panel plots the VQA reliability diagram: the raw-MLE operating point sits on the diagonal at the held-out base rate ($14/18 = 0.778$ of the trials with VQA=yes were truly grasped), confirming that VQA carries no information beyond a constant predictor. The raw-MLE prediction is 0.737 and the Laplace-smoothed prediction (used by the estimator) is 0.747; both are close to the empirical rate. The right panel plots the histogram of post-baseline width residuals from successful held-out trials ($M = 10$); every observation is exactly zero, with the imposed $\mathcal{N}(0, \sigma_w^2)$ overlaid for reference.

5.3 Held-out evaluation

The held-out set comprises 19 trials: 6 slip, 5 missed (all clean-empty), 5 success, and 3 wrong placement, giving 11 positives and 8 negatives for the slip-or-missed detection target. Parameter parity between sessions was verified before scoring: the mean held-grasp width was 0.0324 m on training and 0.0320 m on held-out (standard deviations 1.0 and 0.5 mm respectively), confirming that the physical setup did not drift between collection sessions.

Figure 4 plots NIS per primitive event for a representative held-out slip trial. The signal sits at zero through the close event ($k = 4$) and then spikes to NIS = 441 at the lift ($k = 5$) where the object exits the gripper – three orders of magnitude above the $\chi^2_{0.95}(1) = 3.841$ threshold. The elevated value persists at $k = 6$ because the dropped width has not recovered.

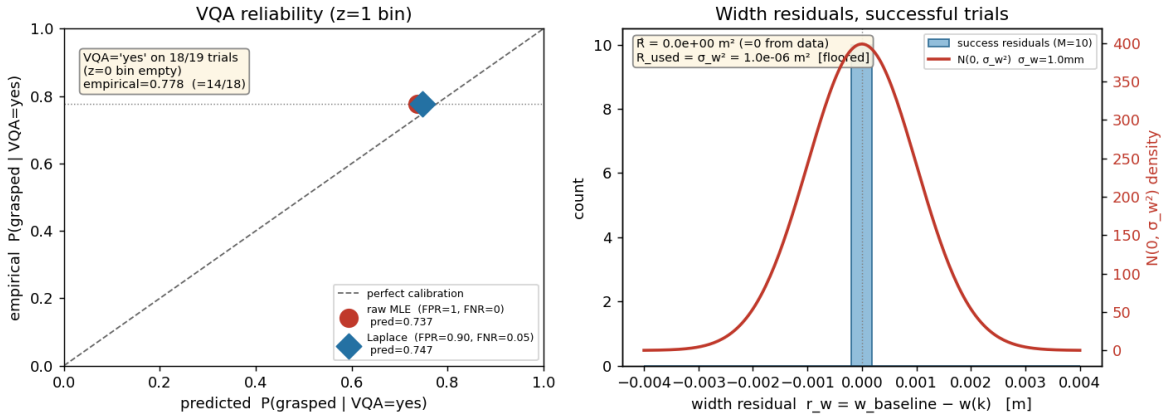


Figure 3: (Left) VQA reliability diagram on the held-out set. The raw-MLE point lies on the diagonal at the held-out base rate (0.778), indicating no information beyond a constant predictor; the Laplace-smoothed point used by the estimator is nearby. (Right) Histogram of post-baseline width residuals from successful held-out trials (M=10); all residuals are exactly zero, with the imposed noise floor $\mathcal{N}(0, \sigma_w^2)$ overlaid.

Method	**Precision**	**Recall**	**TP**	**FP**
Raw VQA ($z_v = 0 \Rightarrow$ failure)	1.000	0.091	1	0
`verify_grasp` (existing heuristic baseline)	1.000	0.455	5	0
NIS tracker at $\chi^2_{0.95}(1) = 3.841$	1.000	0.545	6	0
OR-fusion (NIS tracker $\vee$ `verify_grasp`)	**1.000**	**1.000**	**11**	**0**

Table 3: Held-out detection results on the slip-or-missed target (11 positives, 8 negatives). The OR-fusion of the NIS tracker and the existing heuristic achieves perfect recall at zero false positives.

Table 3 reports the precision and recall of four detectors on the held-out set: the raw VQA answer (predict failure if $z_v = 0$), the existing rule-based `verify_grasp` primitive, the NIS-based width tracker at the chi-squared threshold, and the logical disjunction (OR-fusion) of the latter two. Figure 5 plots the corresponding operating points alongside the full precision-recall curve traced by sweeping the NIS threshold.

5.4 Discussion: sensor complementarity

On training, the NIS tracker caught 6/6 slips plus a single edge-grasp missed trial (the gripper closed on the banana's tip, then lost it during lift), while `verify_grasp` caught the remaining 7 clean-missed trials. On held-out, all 5 missed trials were clean-empty (no edge-grasps occurred), and the two detectors covered exactly disjoint failure subsets: NIS caught all 6 slips and zero missed; `verify_grasp` caught all 5 missed and zero slips. There is a reason for this disjointness: `verify_grasp` is evaluated at the grasp event, before any slip can occur, and therefore cannot detect failures that develop during lift or transit; conversely, the NIS test requires a non-trivial w_b from which width can deviate, which a clean-empty close cannot provide. The two sensors are complementary by design, and the empirical result confirms that their disjunction covers the failure space at zero cost in false-positive rate.

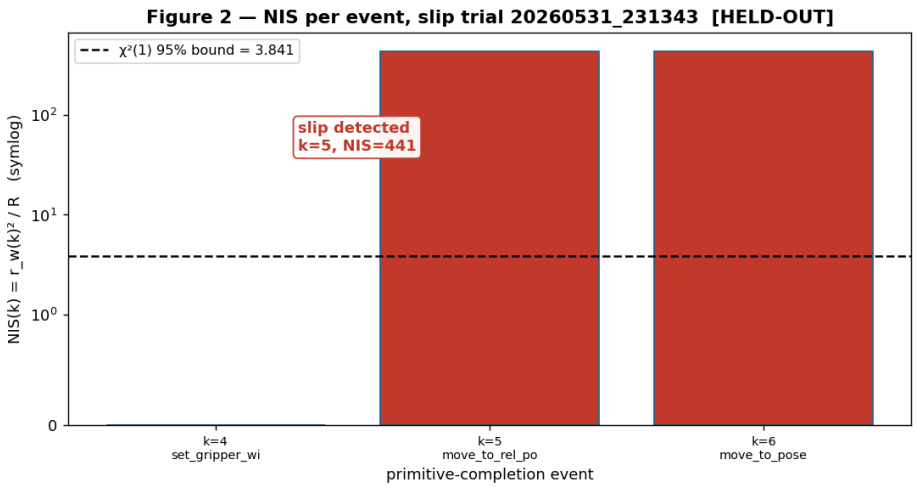


Figure 4: NIS per primitive event for a representative held-out slip trial. The trigger sits silent at the close event, spikes at the lift where the object exits the gripper, and persists at the following move because the width has not recovered. The dashed horizontal line marks the $\chi^2_{0.95}(1) = 3.841$ trigger threshold.

6 Conclusion and Limitations

This work demonstrated that a one-dimensional recursive Bayesian belief tracker, with both sensors calibrated via maximum-likelihood and least-squares system identification and a chi-squared NIS consistency test as the failure trigger, suffices to detect mid-execution slip events on a real Franka manipulator at zero false positives, and that fusing it with the platform's existing rule-based grasp-check yields perfect detection of all grasp failures on a 19-trial held-out set. Two findings from the calibration phase are reported as results, not bugs: the gripper-width encoder is effectively deterministic at primitive-event sampling rates, requiring a physically-motivated noise floor rather than a fitted variance; and the GPT-based VQA sensor is largely non-informative in this configuration, which the Bayes update naturally downweights to almost no contribution.

The principal limitations are of scope and sample size. (i) All experiments use a single task, single object, and single physical configuration; generalization to multi-object or multi-task settings is not demonstrated. (ii) The held-out set contains 11 positive trials, giving a Wilson 95% lower bound on the OR-fusion recall near 0.72 rather than the 1.0 point estimate. (iii) Edge-grasp missed events – in which the gripper closes on the object's edge and loses it during lift, the one regime in which the width tracker catches a "missed" failure – were observed on only one training trial and zero held-out trials; their rate of occurrence in practice is therefore not determined by this study. (iv) The handover trigger was evaluated offline; live integration with the natural-language correction loop is straightforward but was not exercised. (v) The trivial process model $b_g(k \mid k-1) = b_g(k-1)$ omits any belief decay or transition prior, justified here by the event-indexed formulation but a candidate for refinement on tasks with longer or more variable inter-event intervals. Future work includes extending the framework to multiple objects, looking at the edge-grasp case more systematically with a dedicated trial protocol, and wiring the chi-squared trigger into the live correction loop already present in the platform.

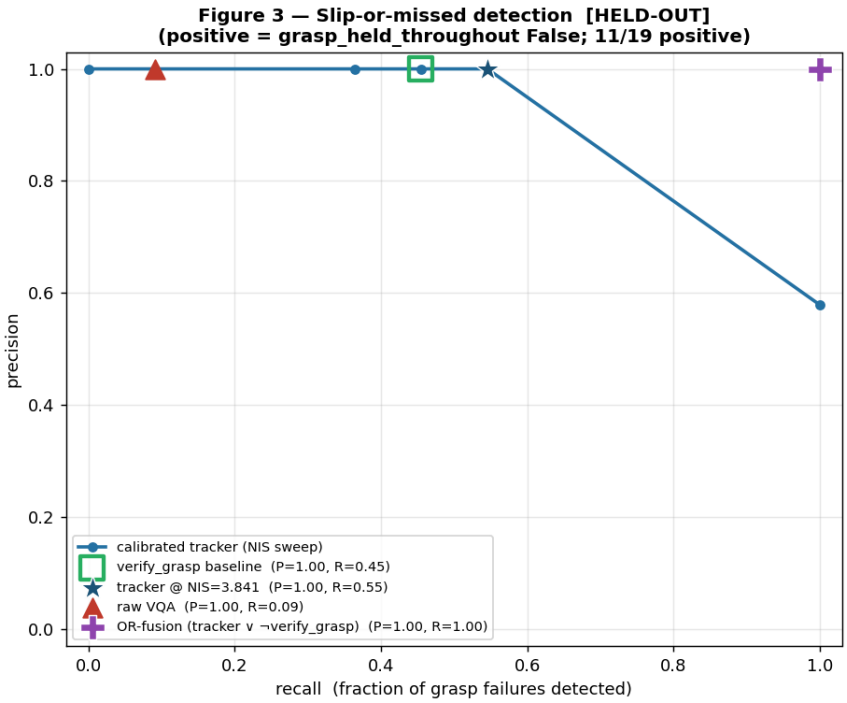


Figure 5: Precision-recall on the held-out set. The NIS tracker (sweep) traces a plateau at precision 1.0 across a wide range of thresholds before falling off. The four operating points of Table 3 are marked; the OR-fusion sits at the top-right corner.

References

[1] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. Wiley, 2001.

[2] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control, 2023. URL `https://arxiv.org/abs/2209.07753`.

[3] M. Murray, A. Gupta, and M. Cakmak. Teaching robots with show and tell: Using foundation models to synthesize robot policies from language and visual demonstration. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4033–4050. PMLR, 06–09 Nov 2025. URL `https://proceedings.mlr.press/v270/murray25a.html`.

[4] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[5] T. Ren et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159, 2024.

[6] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. *IEEE Transactions on Robotics*, 2023.