Before the Needle: Semantic Action Contracts for Risk-Aware Robot Intervention

Mujtaba Hasan
AATUN Labs
New Delhi, India
Email: mujtaba@aatun.in

Ankur Narang
DeepCoreX Inc
Delaware, USA
Email: ankur.narang@deepcorexlabs.com

Abstract*—Robots should not execute irreversible skills merely because a detector is confident. They need an explicit semantic contract that says when perception is sufficient, when another sensing action is useful, and when autonomy must defer. We study this problem through ultrasound-guided needle intervention, where probe motions are reversible information actions but needle insertion can cause harm. We introduce a robot action contract for semantic risk-affordance reasoning. The contract grounds candidate interventions as SAFE, CAUTION, CRITICAL, or UNKNOWN, partitions behavior into *sense*, *act*, and *defer* modes, and uses a policy-induced statistical monitor before irreversible control authority is released. In a feature-level CVC-POMDP, active sensing alone improves coverage but increases unsafe actions, while active sensing plus conformal deferral reduces false-safe interventions from 17.0% to 9.4% relative to passive thresholding. We further add a rendered ultrasound-like perception benchmark. A global image-risk classifier collapses into complete deferral or unsafe low-threshold action, whereas a spatial risk-affordance map reaches 0.784 ± 0.016 mean IoU over five seeds and a Hoeffding-calibrated action contract achieves $71.1 \pm 1.7\%$ act rate, $68.2 \pm 1.4\%$ success, and $2.85 \pm 0.39\%$ false-safe rate on held-out natural episodes. Under close-anatomy shift, the same contract defers on $97.8 \pm 0.7\%$ of episodes. The message is that semantics should be executable: *before a robot commits to an irreversible skill, perception must satisfy an action contract, not just a confidence threshold.

I. INTRODUCTION

A robot system is not only a perception model plus a controller. It is a composition of modules whose outputs must be interpreted before authority is handed to a low-level skill. In many robot stacks, this handoff is implemented through informal confidence thresholds: if the target is visible, run the insertion controller; if the object is detected, grasp it; if the tissue is segmented, cut it. Such interfaces are brittle because the same perceptual statement can mean different things for different downstream actions. "The vessel is visible" may justify another scan; it may not justify inserting a needle.

We study this boundary between perception and control in ultrasound-guided needle intervention. Ultrasound is real-time and portable but noisy, artifact-prone, and deeply embodied: probe translation, rotation, tilt, and pressure all change what the robot sees. Robotic ultrasound systems are increasingly plausible, including autonomous central venous catheterization pipelines [8], adaptive ultrasound-guided needle insertion [11], high-performance robotic ultrasound simulation [3], and anatomy-aware reinforcement learning for intercostal ultrasound scanning [4]. These advances sharpen a system-level question: when should a robot continue sensing, when may it hand control to an irreversible intervention skill, and when must it defer?

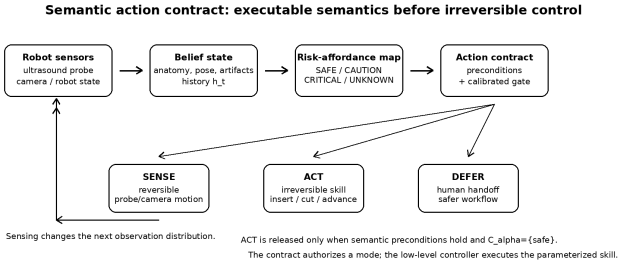

Fig. 1. Robotics-system view. The semantic action contract sits between perception and control. It converts a belief state into risk-affordance semantics and authorizes exactly one mode: reversible sensing, irreversible acting, or deferral. The low-level controller receives an irreversible skill only after semantic preconditions and the calibrated gate pass.

a) Robotics-science question: This paper studies the *action contract*: an explicit interface that turns uncertain perception into authorized robot modes. A sensing skill, or active-sensing action, changes the future observation distribution and is usually reversible. An intervention skill changes the world and may be difficult to undo. A deferral skill transfers authority to a human or safer workflow. The scientific question is therefore not only "what does the robot see?" but "what is the robot permitted to do with what it sees?"

b) Why current-frame confidence is insufficient: A high-confidence ultrasound frame is not necessarily a safe frame. The same local appearance may correspond to a centered target vessel, a partially visible vessel, or a target adjacent to a critical structure. Conversely, a low-confidence frame need not imply immediate handoff if a small probe motion can expose a safer view. This creates three failure modes. A passive policy can act on a misleading current frame. An active policy can find a better view but over-act without calibration. An overly conservative policy can defer even when a reversible sensing action would have made intervention safe.

We propose to make semantics executable. A candidate intervention is grounded as SAFE, CAUTION, CRITICAL, or UNKNOWN. A robot command belongs to one of three modes:

sense, *act*, or *defer*. An irreversible skill is executed only if semantic preconditions hold and a policy-induced conformal monitor returns the singleton set $\{\text{safe}\}$. The resulting system has a clear interpretation: active sensing improves opportunity, calibration controls authorization, and deferral preserves safety when the contract is not satisfied.

Our contributions are:

1) A robotics-system abstraction for irreversible action: semantic action contracts as an interface between perception, planning, low-level control, and supervision.
2) Risk-affordance maps that ground candidate robot interventions into SAFE, CAUTION, CRITICAL, and UNKNOWNregions with explicit preconditions.
3) A policy-induced conformal intervention monitor with finite-sample false-safe control under exchangeability of robot intervention candidates, plus episode-level and shift-aware interpretations.
4) A rendered ultrasound-like perception extension showing that spatial risk-affordance maps are a stronger perception-control interface than global image-risk classification.
5) Closed-loop diagnostics in a CVC-inspired POMDP simulator showing how sensing, acting, deferral, calibration, semantic preconditions, and raw/rendered perception determine robot safety-coverage behavior.

II. Related Work

Semantic affordances and robot action. Semantics become robotics-relevant when they constrain behavior. Affordances, preconditions, skill abstractions, and monitoring rules decide whether a policy should be executed. Surgical affordance work predicts tool-action-specific tissue interaction maps [6], while CholecT50 formalizes instrument-verb-target semantics in endoscopic video [7]. We use a different affordance object: a risk-stratified action contract. A location is not only action-compatible or not; it is safe, cautionary, critical, or unknown for a candidate irreversible skill.

Robotic ultrasound and active perception. Robotic ultrasound aims to reduce operator dependence and improve access to difficult targets. AURA-CVC integrates scanning, vessel localization, insertion-zone selection, and needle guidance [8]. SonoGym provides simulation tasks for ultrasound navigation and surgical guidance [3]. Recent adaptive needle-insertion work studies ultrasound tracking and uncertainty-aware control [11]. Anatomy-aware reinforcement learning for intercostal scanning uses 3D structure to plan around acoustic shadows [4]. Our paper is complementary: rather than proposing a full robotic CVC stack, we isolate the semantic safety layer that arbitrates among sensing, acting, and deferral.

Uncertainty, deferral, and safe control. Active vision studies how agents choose sensing actions to improve downstream decisions [2]. POMDPs formalize planning under partial observability [5]. Safe reinforcement learning and constrained control optimize behavior subject to safety constraints [1]. Conformal prediction gives finite-sample coverage under exchangeability [10, 9]. We use conformal prediction not to produce a better label, but to gate an irreversible robot skill.

System-level robot safety. Our focus is not a new segmentation network, low-level controller, or hardware design. It is the contract layer that decides when existing modules may be composed. This is a robotics-science problem because the same perception output has different consequences depending on the downstream skill: a probe motion updates the belief state, an insertion action commits the robot to a risky trajectory, and deferral changes the locus of authority from autonomy to supervision.

III. Robot Action Contract

We model the robot as a closed-loop system operating over histories $h_t = (o_1, a_1, \ldots, o_t)$ in a partially observable decision process. The latent state is

$$s_t = (z_t, p_t, q_t, n_t, \eta_t), \tag{1}$$

where z_t denotes anatomy or scene geometry, p_t probe or camera pose, q_t robot configuration, n_t tool or needle state, and η_t nuisance factors such as acoustic artifacts, shadows, motion, or occlusion. The observation $o_t \sim O(s_t)$ is a sensory frame, signal, or short history. Low-level skills are assumed to exist: scan adjustment, probe motion, camera motion, insertion, handoff, stop, or tool advancement. The contract decides when each skill is allowed. When the contract authorizes an intervention, it does not directly command motor torques; it releases a parameterized skill such as an insertion point, approach direction, target depth, and allowed motion envelope to a controller that must still enforce hardware-level constraints. This separation is intentional: the paper studies semantic authorization before control authority is released, not the motor controller itself.

The action space is partitioned as

$$\mathcal{A} = \mathcal{A}_{\text{sense}} \cup \mathcal{A}_{\text{act}} \cup \mathcal{A}_{\text{defer}}. \tag{2}$$

Sensing actions $a \in \mathcal{A}_{\text{sense}}$ include probe translation, rotation, tilt, pressure adjustment, camera motion, view change, tissue retraction, or rescan. They are information actions: their primary effect is to change the next observation distribution. Intervention actions $a \in \mathcal{A}_{\text{act}}$ include selecting an insertion point, choosing a needle angle, cutting, clipping, coagulating, or advancing a tool. They are world-changing actions: errors can cause harm. Deferral actions $a \in \mathcal{A}_{\text{defer}}$ hand control to a human, request supervision, or terminate autonomy.

For spatial location x and intervention $a \in \mathcal{A}_{\text{act}}$, define a semantic risk-affordance label

$$y_t(x, a) \in \mathcal{Y} = \{\text{SAFE}, \text{CAUTION}, \text{CRITICAL}, \text{UNKNOWN}\}. \tag{3}$$

A risk-affordance map is the posterior

$$M_t(x, a) = P(y_t(x, a) \mid h_t). \tag{4}$$

The labels are not descriptive tags; they are controller preconditions. SAFEmeans the action reaches a target and avoids known critical structures. CAUTIONmeans the action is close

TABLE I
SEMANTIC ACTION CONTRACTS. THE CONTRACT TURNS PERCEPTION OUTPUTS INTO MODE-LEVEL AUTHORIZATION FOR A ROBOT SKILL.

Mode	Robot meaning	Preconditions / monitor
Sense	reversible information gathering	uncertainty high; predicted reducible; motion cost acceptable
Act	irreversible intervention	target visible; critical margin satisfied; conformal set $\mathcal{C}_\alpha = \{\text{safe}\}$
Defer	handoff or supervision	no certified intervention; sensing unlikely to help or budget exhausted

TABLE II
WHAT CROSSES THE ROBOTICS-SYSTEM BOUNDARY. THE PAPER STUDIES THE CONTRACT LAYER, NOT THE LOW-LEVEL MOTOR CONTROLLER.

Interface	Role in this paper
Perception $\rightarrow$ contract	Convert uncertain sensory belief into risk-affordance semantics
Contract $\rightarrow$ planner	Select sense, act, or defer under semantic preconditions
Contract $\rightarrow$ controller	Release parameterized irreversible skill only after semantic and statistical gates pass
Sensing $\rightarrow$ perception	Change future observation through embodied probe/camera motion
Failure $\rightarrow$ supervision	Defer when the system cannot certify safe autonomy

to a risk boundary. CRITICALmeans the action intersects or threatens a forbidden structure. UNKNOWNmeans the robot lacks visibility or confidence to certify the action. Thus Eq. (3) is not merely a relabeling of perception: it changes the mode-level permission set by making an apparently visible target insufficient for action when critical-margin, uncertainty, or calibration preconditions fail.

The key systems choice is to treat deferral as an action, not as a failure. If the robot cannot satisfy intervention preconditions and cannot predict useful sensing, deferral is the correct semantic behavior. In high-stakes robotics, forcing a choice can convert uncertainty into harm.

a) System invariants: The contract enforces four invariants. First, an irreversible skill cannot be authorized if the calibrated set contains unsafe. Second, a sensing skill is used only when it is expected to reduce action-relevant uncertainty. Third, deferral is always available and terminates autonomous execution. Fourth, calibration is performed on the same policy-induced candidate distribution that the deployed robot will use: candidates generated after active sensing are calibrated against candidates generated by the same sensing policy. These invariants separate semantic authorization from low-level execution.

IV. METHOD

A. Action-conditioned risk-affordance maps

The contract consumes a belief state and produces an authorization decision. The perception side predicts target evidence, critical-structure evidence, visibility, and uncertainty from the current observation and converts them into $M_t(x, a)$. In the CVC abstraction, a planned intervention is safe if it reaches the target vein while maintaining a margin from critical anatomy, caution if it is near a risk boundary, critical if it intersects a forbidden structure, and unknown if visibility or confidence is insufficient.

We evaluate two perception interfaces. The first uses noisy feature-level observations to isolate the contract mechanism over many closed-loop episodes. The second uses rendered ultrasound-like images and learns a pixel-level risk-affordance map with a compact encoder-decoder. The renderer samples vessel geometry, target-critical clearance, depth, visibility, speckle-like noise, shadows, low contrast, and deep-target variants; labels are computed from the same target reachability and critical-margin rules used by the contract. The map model is trained to predict the four risk-affordance classes, while the scalar baseline predicts only whether an image appears globally safe. This raw/rendered perception extension asks whether the robot should consume a scalar image-risk label or a spatial semantic map. The contract layer is agnostic to the encoder: it consumes risk-affordance beliefs and checks whether they satisfy action preconditions. The experiments show, however, that the representation matters operationally. A global image classifier does not produce a useful authorization signal, while spatial maps provide the semantic substrate needed by the contract. To safely utilize these spatial maps, the contract must quantify the uncertainty of the risk-affordance belief before authorizing sensing or action.

B. Action-conditioned sensing

Let $U(M_t)$ be an uncertainty functional such as entropy, ensemble variance, or conformal set size. For each candidate sensing action, the policy predicts future risk-affordance uncertainty,

$$g_\phi(h_t, a) \rightarrow \widehat{U}_{t+1}(a), \qquad a \in \mathcal{A}_{\text{sense}}. \tag{5}$$

It selects a sensing action when expected reduction exceeds cost:

$$\hat{a}_{\text{sense}} = \arg\max_{a \in \mathcal{A}_{\text{sense}}} \Big[U(M_t) - \widehat{U}_{t+1}(a) - \beta c(a) \Big] \tag{6}$$

This local rule has a robotics interpretation: move the sensor only if the movement is predicted to reduce uncertainty about the next irreversible action.

C. Contract monitor and pseudocode

For a candidate intervention $c = (h_t, a_{\text{act}})$, collapse the risk-affordance label into $Y(c) \in \{\text{safe}, \text{unsafe}\}$. Let $f_\theta(y \mid c)$ be a probabilistic safety score and define nonconformity

$$S(c, y) = 1 - f_\theta(y \mid c). \tag{7}$$

On calibration candidates $(c_i, Y_i)_{i=1}^n$, set

$$\hat{q}_\alpha = S_{(\lceil (n+1)(1-\alpha) \rceil)}, \tag{8}$$

TABLE III
ACTION-CONTRACT MONITOR. THE OUTPUT IS A ROBOT MODE, NOT MERELY A CLASS LABEL.

Input: history h_t, sensing skills $\mathcal{A}_{\text{sense}}$, intervention skills $\mathcal{A}_{\text{act}}$, level α
1. Estimate $M_t(x, a)$ and uncertainty $U(M_t)$.
2. For each $a \in \mathcal{A}_{\text{act}}$, build $\mathcal{C}_\alpha(h_t, a)$ from policy-induced calibration.
3. If some $a \in \mathcal{A}_{\text{act}}$ has low risk, low uncertainty, and $\mathcal{C}_\alpha = \{\text{safe}\}$, authorize **Act**.
4. Else estimate $\widehat{U}_{t+1}(a)$ for each $a \in \mathcal{A}_{\text{sense}}$.
5. If the best predicted reduction exceeds cost, authorize **Sense**.
6. Otherwise authorize **Defer**.

where $S_{(k)}$ is the k-th smallest calibration score. The conformal set is

$$\mathcal{C}_\alpha(c) = \{y \in \{\text{safe, unsafe}\} : S(c, y) \leq \hat{q}_\alpha\} \tag{9}$$

The irreversible skill is authorized only if $\mathcal{C}_\alpha(c) = \{\text{safe}\}$.

Calibration must match the policy. Active sensing changes which states reach the intervention gate, so calibrating only on initial observations can evaluate the monitor on an easier distribution than the one encountered after rescanning. We calibrate on candidates induced by the same sensing rule used at test time.

V. SAFETY STATEMENTS

The guarantees below are statistical and scoped. They are not clinical safety guarantees and do not replace robot hardware limits or supervision. They state what the contract monitor controls under explicit assumptions.

Theorem 1 (policy-induced false-safe control). Fix the active sensing rule used before the conformal monitor and let $\mathcal{D}_\pi$ be the distribution over intervention candidates it induces. Assume calibration candidates and a test candidate are exchangeable under $\mathcal{D}_\pi$. If the robot executes an irreversible action only when $\mathcal{C}_\alpha(c) = \{\text{safe}\}$, then

$$P\left(\text{Act} = 1 \ \wedge \ Y(c) = \text{unsafe}\right) \leq \alpha. \tag{10}$$

Proof. Split conformal prediction gives

$$P\left(Y(c) \notin \mathcal{C}_\alpha(c)\right) \leq \alpha. \tag{11}$$

Let $F = \{\text{Act} = 1 \wedge Y(c) = \text{unsafe}\}$ be the false-safe event. If F occurs, then the robot acted. By the decision rule, $\mathcal{C}_\alpha(c) = \{\text{safe}\}$. Since the true label is unsafe, $Y(c) \notin \mathcal{C}_\alpha(c)$. Therefore

$$F \subseteq \{Y(c) \notin \mathcal{C}_\alpha(c)\}. \tag{12}$$

Taking probabilities yields

$$P(F) \leq P(Y(c) \notin \mathcal{C}_\alpha(c)) \leq \alpha. \quad \square \tag{13}$$

This controls $P(\text{Act} = 1, Y = \text{unsafe})$, not $P(Y = \text{unsafe} \mid \text{Act} = 1)$. The distinction is important. A cautious robot may act rarely, so conditional unsafe rate can remain high even when episode-level false-safe events are reduced. We therefore report both action rate and false-safe rate per episode.

Corollary 1 (episode-level union control). If an episode proposes at most K intervention candidates and each candidate uses level α/K, then the probability of at least one false-safe intervention in the episode is at most α, by a union bound over candidate false-safe events. This conservative rule turns a candidate monitor into an episode-level operating policy when a hard candidate budget is known.

Proposition 1 (one-step value of information). Let $\bar{R}(h, a)$ be an upper confidence bound on intervention risk. A sensing action a_s is preferred to immediate intervention under a one-step risk proxy when

$$\mathbb{E}\left[\min_{a \in \mathcal{A}_{\text{act}}} \bar{R}(h_{t+1}, a) \mid h_t, a_s\right] + \beta c(a_s) < \min_{a \in \mathcal{A}_{\text{act}}} \bar{R}(h_t, a). \tag{14}$$

This follows by comparing the cost-adjusted post-sensing risk bound to the current best intervention-risk bound. The proposition is local, not globally POMDP-optimal. Its role is to justify sensing only when a robot expects the next observation to reduce intervention risk.

Shift remark. If the false-safe event has probability at most α under the calibration candidate distribution P and the deployed candidate distribution is Q, then

$$Q(F) \leq \alpha + \text{TV}(P, Q). \tag{15}$$

Here $\text{TV}(P, Q)$ is total variation distance. The bound follows directly from its definition. It motivates policy-induced calibration and stress testing: when sensing changes the distribution of candidates, the calibration data must change with it.

VI. EXPERIMENTAL SYSTEM

A. Simulator and design goal

We evaluate the contract as a closed-loop robot system in a lightweight central-venous-catheterization-inspired POMDP simulator. Each episode samples a target vein, a nearby critical structure, vessel radii, probe pose, artifact level, and candidate insertion point in a 2D planned-insertion abstraction. The robot observes noisy target and critical estimates with visibility and uncertainty features. Sensing actions move the probe and update the observation. Intervention commits to the current candidate point. Deferral terminates without intervention.

The simulator is not intended to be a clinical ultrasound model. It abstracts away raw image formation, 3D tissue, contact force, needle bending, patient motion, and robot dynamics. Its purpose is to isolate a system mechanism: how an embodied robot arbitrates among reversible sensing, irreversible action, and supervised handoff when perception is uncertain.

B. Baselines, diagnostics, and metrics

We compare passive thresholding, random rescan, entropy-greedy sensing, conformal-only deferral, active sensing without conformal filtering, active sensing with policy-induced conformal filtering, oracle uncertainty sensing, and a safe oracle. Passive thresholding uses no probe motion and acts on the current candidate if a risk score crosses a fixed threshold. Active sensing means a reversible information action changes

TABLE IV
CVC-POMDP DESIGN CHOICES. THE BENCHMARK ISOLATES THE CONTRACT LAYER RATHER THAN CLINICAL REALISM.

Component	Simulator abstraction
Latent anatomy	vein, artery, radii, depth, artery-vein distance
Observation	noisy target/critical estimates; visibility; uncertainty; rendered US-like snapshots for inspection
Sensing actions	probe shifts in eight directions with scan cost
Intervention	insert at estimated target-vessel center
Risk labels	safe, caution, critical, unknown via target and margin rules
Shifts	speckle, shadow, low contrast, close anatomy, deep target, biased perception

TABLE V
EVALUATION LOGIC: EACH DIAGNOSTIC PROBES A ROBOTICS-SYSTEMS PROPERTY OF THE ACTION CONTRACT RATHER THAN A STATIC PERCEPTION SCORE.

Diagnostic	Robotics question
Main baselines	Does active sensing improve safety or only coverage?
Semantic ablation	Which preconditions actually prevent unsafe action?
Alpha sweep	Can the contract expose a tunable safety-coverage knob?
Calibration mismatch	Does the monitor depend on the policy-induced candidate distribution?
Rendered perception	Does a spatial risk-affordance map improve the perception-control interface?
Stress tests	Does the robot become conservative under degraded perception?
Qualitative trajectories	Are failures traceable to action-mode decisions?

the next observation before intervention authorization. "Ours" combines semantic active sensing with reducibility and margin preconditions; "Ours + conformal" adds the policy-induced conformal singleton gate. Oracle uncertainty chooses the best information action but is not a safety oracle, while the safe oracle acts only on ground-truth safe candidates. Additional diagnostics test semantic precondition ablations, calibration mismatch, alpha sweeps, and stress settings.

The primary safety metric is false-safe intervention rate per episode: how often the robot commits to an intervention that is unsafe under simulator ground truth. Success is a safe intervention on the target; critical intervention is the subset of unsafe interventions that threaten the critical structure; action rate is the probability of releasing the intervention skill; deferral is the probability of handing off or terminating autonomy; and rescans count reversible information actions. Expected calibration error (ECE) is the binned absolute gap between predicted safety confidence and empirical safety frequency on intervention candidates. Unsafe $\mid$ act is the conditional unsafe rate among episodes in which the robot acted, which is reported separately from episode-level false-safe rate because a very conservative robot may act rarely. The action-conditioned diagnostic reports top-1 agreement with the oracle best sensing action, regret relative to oracle uncertainty sensing, and correlation between predicted and realized future uncertainty.

The evaluation is therefore not a standard segmentation or tracking benchmark. It is a closed-loop contract benchmark. The unit of analysis is an episode in which the robot must decide whether to change viewpoint, execute an intervention skill, or transfer authority. A method succeeds only if it improves useful intervention without converting uncertainty into false-safe action. For reproducibility, reported means use the same five seeds across policies, disjoint calibration and held-out test episodes, and policy-induced calibration candidates generated by the deployed sensing rule. The rendered pipeline uses calibration only for threshold selection; held-out natural and stress suites never choose the threshold.

VII. RESULTS

The results are organized as robotics system diagnostics. We ask: Does active sensing alone make the robot safer? Does the contract layer change authorization behavior? Does deferral prevent forced unsafe action? Does the conformal monitor expose a tunable safety-coverage interface?

A. Active sensing needs a semantic safety monitor

Table VI shows the main multi-seed results. Active sensing alone increases coverage but also increases unsafe action. Ours without conformal filtering improves success from 17.1% to 25.6% relative to passive thresholding, but false-safe interventions rise from 17.0% to 29.4%. This is the central negative result: better sensing is not the same as safer action.

Adding policy-induced conformal filtering changes the operating point. Ours + conformal improves success relative to conformal-only deferral (20.2% vs. 11.6%), while reducing false-safe interventions relative to passive thresholding (9.4% vs. 17.0%). This is the desired semantic behavior: gather information when it helps, but do not let more information create more unsafe interventions.

B. Rendered perception: scalar labels versus spatial maps

The preceding diagnostics use feature-level observations to isolate the contract mechanism. We next test the perception-control interface on rendered ultrasound-like images. The question is not whether the renderer is clinically realistic; it is whether a robot should authorize action from a scalar image label or from a spatial semantic map. A global image-risk classifier fails as an interface: at the default threshold it defers on all balanced cases, and threshold sweeping gives only 1.25% success with 87.5% conditional unsafe rate among acted cases. In contrast, a pixel-level risk-affordance model learns safe, caution, critical, and unknown regions. Over five seeds it obtains $98.38 \pm 0.19\%$ pixel accuracy, safe IoU 0.624 ± 0.030, critical IoU 0.906 ± 0.008, unknown IoU 0.879 ± 0.009, and mean IoU 0.784 ± 0.016.

TABLE VI

MAIN CVC-POMDP RESULTS. VALUES ARE PERCENTAGES EXCEPT RESCANS. MEAN ± STANDARD ERROR OVER FIVE SEEDS. ACTIVE SENSING WITHOUT A CALIBRATED SEMANTIC MONITOR INCREASES UNSAFE ACTION; ACTIVE SENSING PLUS CONFORMAL DEFERRAL IMPROVES THE SAFETY-COVERAGE TRADEOFF.

Policy	Success ↑	False-safe ↓	Critical ↓	Act	Defer / fallback	Rescans
Passive threshold	17.1±0.5	17.0±0.9	1.2±0.3	34.1±1.0	65.9±1.0	0.00±0.00
Random rescan	21.3±0.4	25.5±0.8	1.7±0.4	46.8±0.9	53.2±0.9	1.83±0.03
Entropy-greedy	30.2±0.9	33.6±0.7	2.7±0.2	63.8±1.2	36.2±1.2	0.82±0.03
Conformal-only	11.6±0.7	7.0±0.7	0.3±0.1	18.6±0.8	81.4±0.8	0.00±0.00
Ours	25.6±0.7	29.4±1.3	2.2±0.3	55.0±0.9	45.0±0.9	0.66±0.02
Ours + conformal	**20.2±0.5**	**9.4±0.6**	**0.7±0.2**	**29.5±0.7**	**70.5±0.7**	**0.88±0.02**
Oracle uncertainty	37.1±1.0	45.6±1.2	2.6±0.1	82.8±0.6	17.2±0.6	1.03±0.01
Safe oracle	70.2±0.6	0.0±0.0	0.0±0.0	70.2±0.6	29.8±0.6	1.09±0.03

TABLE VII

RENDERED ULTRASOUND-LIKE PERCEPTION AND ACTION-CONTRACT RESULTS. VALUES ARE FRACTIONS, MEAN ± STANDARD ERROR OVER FIVE SEEDS. A SCALAR IMAGE CLASSIFIER IS EITHER OVER-CONSERVATIVE OR UNSAFE AT LOW THRESHOLDS; A SPATIAL RISK-AFFORDANCE MAP SUPPORTS USEFUL CALIBRATED ROBOT AUTHORIZATION.

Method	Act rate ↑	Success ↑	False-safe ↓	Defer	Unsafe | act ↓
Global image classifier	0.000	0.000	0.000	1.000	0.000
Global classifier, swept	0.100	0.0125	0.0875	0.900	0.875
Risk-map contract	0.750±0.012	0.724±0.009	0.0255±0.0053	0.250±0.012	0.0337±0.0067
Risk-map, swept	0.832±0.0065	0.739±0.0075	0.0925±0.0033	0.168±0.0065	0.111±0.0041
Risk-map + Hoeffding calibration	**0.711±0.017**	**0.682±0.014**	**0.0285±0.0039**	**0.290±0.017**	**0.0399±0.0049**

Table VII is the strongest evidence for the semantic-contract thesis. The useful perception output is not a single confidence score; it is a spatial map that says where action is safe, cautionary, critical, or unknown. We then select the action threshold using a Hoeffding upper confidence bound on calibration false-safe rate. On held-out natural episodes, this yields $71.1 \pm 1.7\%$ act rate, $68.2 \pm 1.4\%$ success, and $2.85 \pm 0.39\%$ false-safe rate.

C. Semantic preconditions are not cosmetic

Table VIII tests whether semantic action contracts matter. Removing the conformal precondition increases success but more than doubles false-safe rate. Removing deferral entirely forces the robot to act and produces the highest false-safe burden. Removing the critical-margin or reducibility precondition also makes the policy less conservative. The columns are outcome rates, not a mutually exclusive partition: critical interventions are included within false-safe interventions, and rounding can move success, false-safe, and defer slightly away from 100. This supports the SemRob thesis: semantic structure matters when it changes whether a robot is allowed to act.

D. Risk-coverage and calibration diagnostics

Figure 2 sweeps the intervention threshold, i.e., the minimum predicted safety score required before an intervention candidate can be considered for authorization. Each point is one operating setting. Conservative settings move toward high deferral and low false-safe rate; permissive settings move toward higher coverage and higher false-safe rate. The specific insight is that active sensing shifts the robot toward more opportunities to act, but only the calibrated contract moves the operating point toward lower false-safe rate at comparable deferral.

TABLE VIII

SEMANTIC PRECONDITION ABLATIONS IN AN ADDITIONAL DIAGNOSTIC SETTING. VALUES ARE PERCENTAGES. THIS ABLATION USES A SEPARATE DIAGNOSTIC CONFIGURATION TO ISOLATE SEMANTIC PRECONDITIONS AND IS NOT INTENDED TO NUMERICALLY MATCH THE MAIN MULTI-SEED OPERATING POINT IN TABLE VI. COLUMNS ARE NOT MUTUALLY EXCLUSIVE BECAUSE CRITICAL INTERVENTIONS ARE A SUBSET OF FALSE-SAFE INTERVENTIONS. REMOVING THE CONFORMAL PRECONDITION OR DEFERRAL TURNS UNCERTAINTY INTO UNSAFE ACTION.

Variant	Success ↑	False-safe ↓	Critical ↓	Defer
Full semantic + conformal	25.9	17.7	0.7	56.3
No conformal precondition	35.5	40.3	1.7	24.3
No critical-margin precondition	26.1	20.5	1.5	53.5
No reducibility precondition	27.2	20.7	0.7	52.1
No deferral / forced action	47.1	52.9	3.1	0.0

Figure 3 provides a second view of the safety-coverage interface by sweeping the conformal level α. At $\alpha = 0.01$, success is 10.0% and false-safe rate is 4.0% with 86.0% deferral. At $\alpha = 0.20$, success rises to 38.8%, but false-safe rate rises to 36.2% and deferral falls to 25.0%. The contract exposes a knob that a deployment process could set according to task risk and supervision availability.

For the rendered perception pipeline, we also select thresholds with a finite-sample upper confidence bound: $\widehat{R}_{\text{false-safe}} + \sqrt{\log(1/\delta)/(2n)} \leq \alpha$ with $\alpha = 0.10$, $\delta = 0.05$, and $n = 400$. Across five seeds, the selected thresholds average 0.52 ± 0.08. The held-out test false-safe rate is $2.85 \pm 0.39\%$, below the target despite threshold variation across seeds. This is a conservative validation rule rather than a clinical guarantee; its role is to prevent optimistic threshold selection from turning a map predictor into an unsafe actuator.

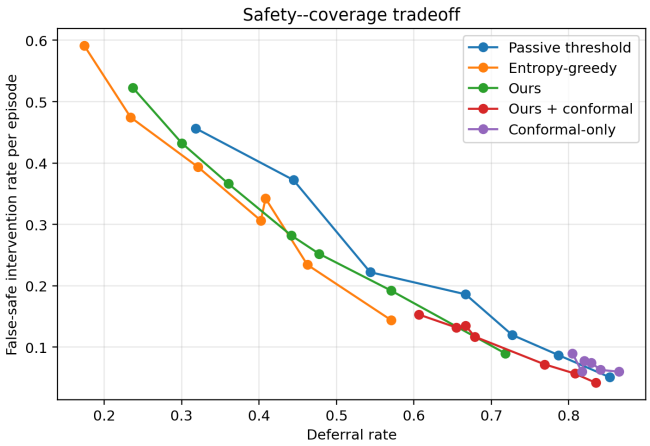

Fig. 2. False-safe intervention rate versus deferral rate across thresholds. Active sensing improves coverage, but calibrated deferral is needed to convert sensing into a safer operating regime.

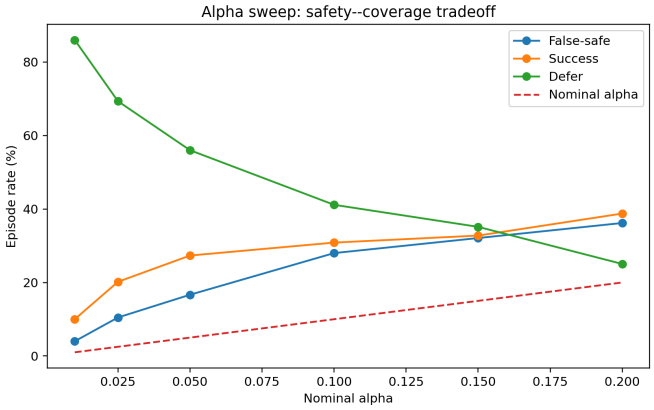

Fig. 3. Alpha sweep. Lower conformal levels increase deferral and reduce false-safe interventions; higher levels authorize more actions but accept greater risk.

E. Action-conditioned uncertainty and closed-loop behavior

The action-conditioning diagnostic shows that predicted future uncertainty is meaningful but imperfect: mean correlation with oracle future uncertainty is 0.815, top-1 agreement with the oracle best information action is 54.8%, and mean regret is 0.022 with median regret 0.0. This supports the mechanism behind sensing while leaving room for learned action-conditioned predictors.

Figure 4 illustrates three recurring patterns in the cleaned camera-ready schematic. Successful episodes start with high uncertainty and use one or two probe motions before the conformal set becomes the singleton safe label. Some active-only failures are prevented because improved target visibility does not remove uncertainty about critical-structure proximity. In severe artifact cases, recursive rescanning is prevented by the reducibility and budget preconditions: if predicted uncertainty reduction falls below cost or the budget is exhausted, the contract selects defer rather than continuing to sense.

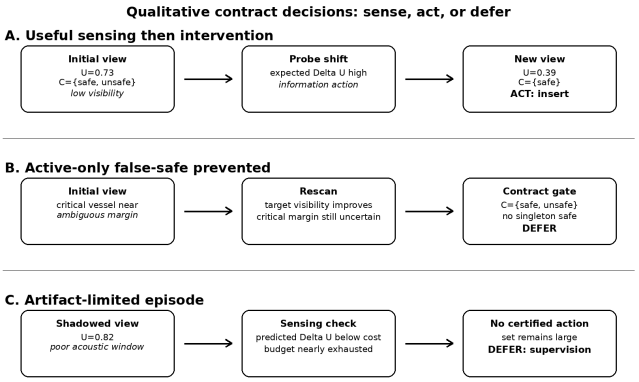

Fig. 4. Qualitative decision trajectories. The cleaned schematic separates three cases: useful sensing followed by insertion, an active-only false-safe prevented by the conformal set, and an artifact-limited episode where reducibility fails and the robot defers.

TABLE IX
RENDERED PERCEPTION STRESS TESTS WITH HOEFFDING-CALIBRATED THRESHOLDS. VALUES ARE FRACTIONS, MEAN $\pm$ STANDARD ERROR OVER FIVE SEEDS. CLOSE ANATOMY CORRECTLY TRIGGERS NEAR-COMPLETE DEFERRAL.

Suite	Act rate	Success	False-safe	Defer
Natural test	0.711 ± 0.017	0.682 ± 0.014	0.0285 ± 0.0039	0.290 ± 0.017
Speckle	0.725 ± 0.010	0.683 ± 0.010	0.0415 ± 0.0027	0.276 ± 0.010
Shadow	0.669 ± 0.021	0.634 ± 0.020	0.0345 ± 0.0041	0.332 ± 0.021
Low contrast	0.709 ± 0.026	0.677 ± 0.023	0.0320 ± 0.0041	0.292 ± 0.026
Deep target	0.678 ± 0.020	0.641 ± 0.015	0.0370 ± 0.0060	0.323 ± 0.020
Close anatomy	0.0225 ± 0.0066	0.0085 ± 0.0027	0.0140 ± 0.0039	0.978 ± 0.0066

F. Stress tests

The feature-level contract is conservative under held-out imaging and anatomy shifts, reducing average false-safe rate from 16.5% to 8.1% and critical interventions from 1.65% to 0.75% relative to passive thresholding. The rendered perception pipeline shows the same qualitative behavior with learned maps. Table IX applies the Hoeffding-calibrated threshold selected on natural calibration episodes to held-out stress suites. Speckle, shadow, low contrast, and deep target shifts preserve high automation coverage while keeping false-safe rates near 3–4%. The close-anatomy shift is different: critical anatomy is intentionally near the target, so the contract almost always refuses to act, deferring on $97.8 \pm 0.7\%$ of episodes. This is the intended semantic failure mode for irreversible action.

VIII. ROBOTICS-SYSTEM INTERPRETATION

The main result is a decomposition of autonomy authority. Active sensing by itself is a coverage mechanism: it creates more chances to act. Conformal deferral by itself is a safety mechanism: it rejects many candidates, including some useful ones. The action contract is useful because it composes the two. It asks the robot to sense when information is reducible, act only when the semantic and statistical gates pass, and defer when the contract remains unsatisfied.

This interpretation is closer to robot systems than to static perception. A detector answers whether a structure is present. A controller answers how to execute a trajectory. The contract answers whether the trajectory should be made available to the controller at all. This middle layer is where semantic robotics can have operational force: semantic states become preconditions, uncertainty becomes a mode switch, and deferral becomes an action in the robot policy.

The contract also clarifies accountability. If the robot acts unsafely, the failure can be localized: the perception belief was wrong, the risk-affordance semantics were wrong, the sensing-value estimate was wrong, the conformal monitor was miscalibrated, or the low-level controller violated the authorized skill. This is a systems advantage over a monolithic policy whose action cannot be decomposed into perception, semantic authorization, and motor execution.

The same structure applies beyond ultrasound. In surgical robotics, camera motion or tissue retraction can gather information, while cutting, clipping, coagulation, or forceful grasping are hard-to-undo actions. In lab robotics, imaging and verification can precede pipetting, mixing, or sample transfer. In field robotics, viewpoint changes can precede irreversible entry into hazardous regions. The concrete labels differ, but the contract pattern remains: sense if information can help, act only if preconditions are certified, and defer otherwise.

IX. Limitations and Path to a Full Robot System

The simulator does not solve ultrasound intervention. It abstracts away real ultrasound formation, 3D anatomy, contact mechanics, patient motion, robot dynamics, needle deformation, and human workflow. The latest version adds rendered ultrasound-like images and learned risk-affordance maps, but these are still synthetic images, not real probe logs or phantom data. The principal limitation is therefore no longer merely the absence of an image pathway; it is the absence of physically validated ultrasound image formation and robot action logs.

A stronger robot system would add four layers. First, real raw perception: train risk-affordance maps on ultrasound frames, robot pose, and anatomical priors rather than rendered images alone. Second, action-conditioned dynamics: train g_ϕ from probe-action transitions or high-fidelity simulation. Third, physical safety: combine the semantic contract with force limits, motion limits, sterile-zone constraints, inverse-kinematics feasibility, contact-force monitoring, and controller-level shields. Fourth, prospective validation: evaluate on phantom systems, logged probe trajectories, and clinician-defined failure modes.

Anatomy-aware planning methods and robotic ultrasound simulators provide a plausible path. 3D priors can make action-conditioned sensing more realistic under acoustic access constraints [4], and simulation platforms such as SonoGym can supply richer observation-action transitions [3]. Dense surgical affordance prediction [6] supplies spatial actionability; the contribution here is to make such actionability risk-stratified, policy-induced, and connected to a gate before irreversible control.

The conformal guarantee is limited. It is marginal, candidate-level, and depends on exchangeability of policy-induced candidates. It does not control all harms, conditional unsafe rate given action, hardware failures, arbitrary distribution shift, or clinical outcomes. These limitations are not incidental. They define the boundary between a statistical monitor and a deployed medical robot. A real system must combine the contract with physical constraints, out-of-distribution detection, clinician oversight, and task-specific validation.

X. Conclusion

We introduced semantic action contracts for risk-aware robot intervention. The framework separates reversible sensing, irreversible acting, and deferral; grounds candidate skills into risk-affordance labels; and uses a policy-induced statistical monitor to block false-safe authorization. In a CVC-inspired simulator, active sensing alone improved coverage but increased unsafe actions, while calibrated semantic deferral produced a safer safety-coverage tradeoff. In rendered ultrasound-like perception, spatial risk-affordance maps provided a much stronger robot-action interface than global image labels and triggered conservative fallback under close anatomy. Before the needle, the robot should not merely perceive an opportunity. It should prove that the action contract permits control.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31, 2017.

[2] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

[3] Yunke Ao, Masoud Moghani, Mayank Mittal, Manish Prajapat, Luohong Wu, Frederic Giraud, Fabio Carrillo, Andreas Krause, and Philipp Fürnstahl. SonoGym: High performance simulation for challenging surgical tasks with robotic ultrasound. *arXiv preprint arXiv:2507.01152*, 2025.

[4] Yuan Bi, Cheng Qian, Zhicheng Zhang, Nassir Navab, Zhongliang Jiang, et al. Autonomous path planning for intercostal robotic ultrasound imaging using reinforcement learning. *Scientific Reports*, 16:6356, 2026. doi: 10.1038/s41598-026-37702-9.

[5] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

[6] Aiza Maksutova, Lalithkumar Seenivasan, Hao Ding, Jiru Xu, Chenhao Yu, Chenyan Jing, Yiqing Shen, and Mathias Unberath. AffordTissue: Dense affordance prediction for tool-action specific tissue interaction. *arXiv preprint arXiv:2604.01371*, 2026.

[7] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. doi: 10.1016/j.media.2022.102433.

[8] Deepak Raina, Lidia Al-Zogbi, Brian Teixeira, Vivek Singh, Ankur Kapoor, Thorsten Fleiter, Muyinatu A. Lediju Bell, Vinciya Pandian, and Axel Krieger. AURA-CVC: Autonomous ultrasound-guided robotic assistance for central venous catheterization. *International Journal of Computer Assisted Radiology and Surgery*, 2026. doi: 10.1007/s11548-026-03572-9.

[9] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[10] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[11] Yuelin Zhang, Qingpeng Ding, Longxiang Tang, Chengyu Fang, and Shing Shin Cheng. A vision-language-action model for adaptive

ultrasound-guided needle insertion and needle tracking. *arXiv preprint arXiv:2604.20347*, 2026.