C-VOIR: TASK INFERENCE FOR HETEROGENEOUS ROBOT TEAMS

Anonymous authors
Paper under double-blind review

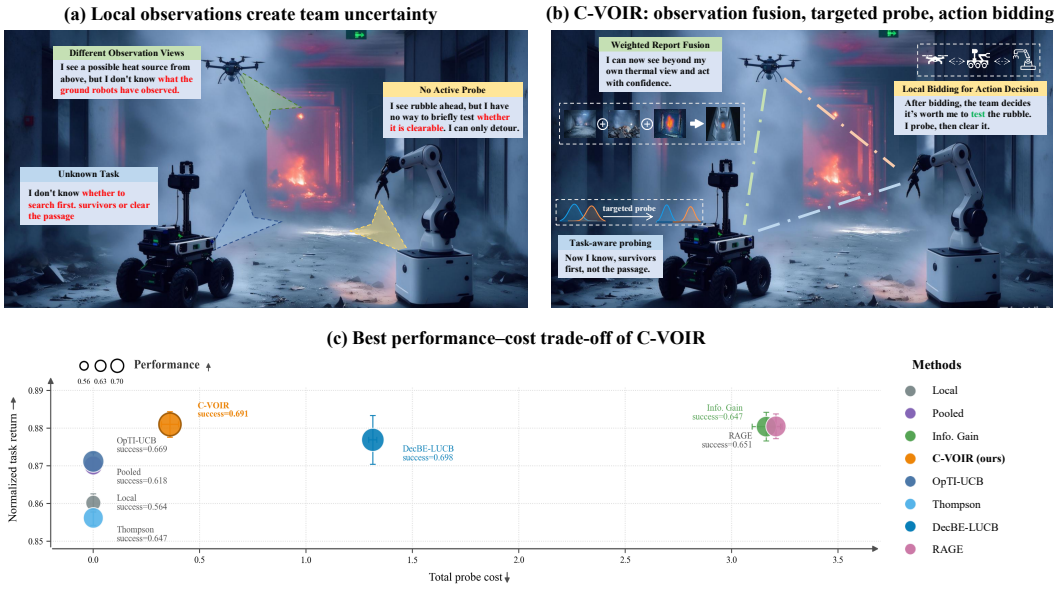

Figure 1: C-VOIR addresses heterogeneous task inference through quality-aware evidence fusion, decision-relevant probing, and cost-aware team agreement. Panel (c) shows the resulting performance–cost–success trade-off on HeteroArm-Mission.

ABSTRACT

Behavior foundation models (BFMs) provide reusable policies, but an unknown deployment reward requires evidence before a policy can be selected. We study online task inference in heterogeneous robot teams whose members have different embodiments, sensors, actions, and probe costs, and communicate only with neighbors. We propose C-VOIR, a cost-aware value-of-information method that weights reports by sensor noise, age, and prediction consistency. It values physical probes by their expected reduction of uncertainty between competing decision labels minus physical cost. Robots exchange their best positive bids, and a strict-majority lease authorizes at most one probe. The team executes a common policy label only when fresh, consistent opinions indicate that no worthwhile probe remains and the decision gap supports a winner; otherwise, it retains baseline actions and no-op probes. Across three evaluation trials, C-VOIR achieves the highest reported epoch-10 running-average return on four open-sourced benchmarks. We also introduce HeteroArm-Mission, a 600-instance MuJoCo benchmark with two- and three-joint arms, heterogeneous observations, and physical probes. C-VOIR achieves the highest reported non-Oracle task return 0.8810 and cost-adjusted return 0.8750, while reducing mean probe cost by 88.6% relative to Information Gain. Our code is open-sourced at https://anonymous.4open.science/r/Anonymous_code-77E6.

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.

1 INTRODUCTION

A behavior foundation model (BFM) provides reusable policies, but **the reward that determines which policy is appropriate may be unknown at deployment**. For example, in a fire-rescue mission, a robot team may have policies for inspecting a possible survivor, opening a blocked passage, and limiting disturbance, yet not know which objective should be prioritized in the current situation. The team must infer these task preferences from interaction feedback before selecting a policy. We study this problem in heterogeneous robot teams whose members differ in embodiment, sensing, and native actions. **Each robot has only local evidence and communicates with its neighbors. The team must combine these observations, decide whether a physical probe is worth its cost, and determine when the evidence is sufficient to support a common policy decision.**

OpTI-BFM infers an unknown linear reward through online interaction, but assumes a single robot and decision maker Rupf et al. (2026). Heterogeneous information-gathering methods address energy, motion, and communication constraints Cai et al. (2023), while calibrated multi-robot planning considers when to request help Wang et al. (2025). Neither setting jointly addresses online policy selection, decision-relevant physical probes, and team agreement from different local evidence.

Three difficulties are central. Reports vary in sensor noise, freshness, and consistency, and a forward report must not be counted twice or stale evidence will dominate the team belief. A probe can reduce uncertainty without changing the policy decision, yet incurs time, energy, disturbance, and recovery cost, without a stopping rule, probing continues indefinitely. Finally, robots with different posteriors must authorize at most one optional probe and avoid executing an unsupported common decision, without fusion, redundant probes and inconsistent actions follow.

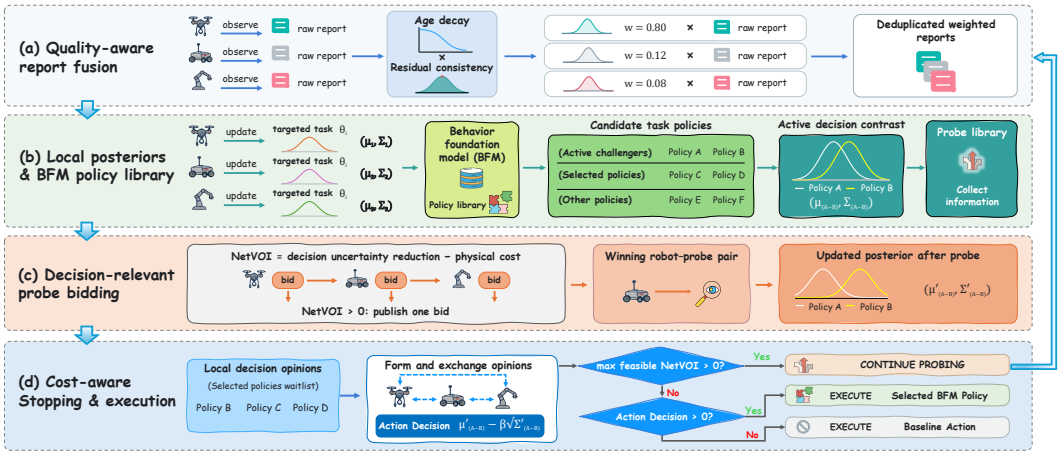

Figure 2: Overview of C-VOIR. (a) *Fusion* deduplicates and weights exchanged reports to align local task beliefs. (b) Each robot identifies uncertain choices from its posterior and BFM policy library. (c) A *probe* explores information that may change the decision; a *bid* proposes each robot's best probe, and a majority-backed *lease* authorizes one winning robot–probe pair. (d) *Stopping* ends probing when consistent opinions report no positive-value probe, after which the team executes the supported policy or retains its baseline actions.

We propose Cost-aware Value-of-Information Reasoning (C-VOIR), illustrated in Figure 2. A BFM supplies candidate policies, while each robot maintains a posterior over the shared unknown reward. Common decision labels allow different embodiments to agree on a policy choice while executing their own native actions. C-VOIR weights uniquely identified reports according to declared sensor noise, age, and consistency with the current prediction. It values each feasible probe by its expected reduction of uncertainty between competing decision labels minus its physical cost. Each robot then communicates at most one proposal—its feasible probe with the largest positive net value, together with the robot identity, probe identity, estimated value, and cost. We call this proposal a *bid*; it proposes a probe but does not by itself authorize execution. Robots propagate and deterministically rank the received bids. A probe may be executed only after a strict majority confirms the same winning robot–probe pair. We call this temporary majority-backed authorization a *lease*; only the named

robot may execute the named probe during the current decision epoch, and missing or conflicting confirmations result in no optional probe. Finally, task execution requires fresh and consistent opinions, no worthwhile remaining probe, and a positive decision gap. Otherwise, the robots retain their baseline actions and no-op probes.

We evaluate C-VOIR over three trials in two settings. Walker, Cheetah, and Quadruped from Ex-ORL/DMC test online task inference on established control domains, but share one embodiment and behavior library. We therefore introduce HeteroArm-Mission, a 600-instance MuJoCo benchmark with two- and three-joint arms, different action libraries and observation noise, and physical probes across four task families. C-VOIR has the highest reported epoch-10 running-average return in all three DMC domains. On HeteroArm-Mission, it obtains the highest non-Oracle task return 0.8810 and net return 0.8750, with 88.6% lower mean probe cost than Information Gain. Adapted DecBE-LUCB has slightly higher team ϵ-success, so the advantage is a performance–cost trade-off, not dominance on every metric.

Our contributions are:

- We formulate online task inference for heterogeneous robot teams and introduce HeteroArm-Mission to evaluate embodiment-specific actions, noisy local observations, and physical probe costs.
- We develop report-level evidence weighting based on sensor noise, age, and prediction consistency, while preventing duplicate reports from adding evidence.
- We define decision-relevant, cost-adjusted probe value and coordinate one positive-value probe through bids and a strict-majority lease.
- We give an opinion-based stopping rule that authorizes task execution only when no worthwhile probe remains and the common decision has a positive gap; otherwise, the team retains its baseline.

2 Related Work

2.1 Behavior models and decision-focused information gathering

Zero-shot reinforcement learning trains reusable behavior models before the downstream reward is known. Forward–Backward representations select policies from test-time rewards Touati et al. (2023), while subsequent work improves imitation, task representations, offline training, and transfer Pirotta et al. (2024); Park et al. (2024); Frans et al. (2024); Ingebrand et al. (2024); Tirinzoni et al. (2025). Other methods improve behavior-space coverage, representations, and online adaptation Agarwal et al. (2025); Cetin et al. (2025); Sun et al. (2025); Sikchi et al. (2025); Jajoo et al. (2026). OpTI-BFM is closest to our setting: it infers unknown linear reward weights through online interaction, but assumes a single robot and decision maker Rupf et al. (2026). Decision-focused information gathering selects observations for a downstream choice. IDRL reduces uncertainty between plausible policies, while RAGE separates the measurement set from the final decision set Lindner et al. (2021); Fiez et al. (2019). Later work considers observation cost, query–policy mismatch, and measurement fidelity Tucker et al. (2023); Hu et al. (2024); Poiani et al. (2024). Unlike these methods, C-VOIR selects embodiment-specific physical probes according to their expected effect on a shared team decision and their physical cost.

2.2 Distributed learning and multi-robot coordination

Distributed bandits study shared-parameter learning through neighbor consensus and communication-limited exchange Amani et al. (2023); Salgia & Zhao (2023); Mitra et al. (2023), including asynchronous communication and heterogeneous rewards Chen et al. (2023); Richard et al. (2024); Wang et al. (2023). Multi-robot methods coordinate sensing, localization, assignment, and exploration under heterogeneous capabilities Cai et al. (2023); Spasojevic et al. (2023); Goarin & Loianno (2024); Tian et al. (2024), as well as connectivity, language-based planning, and active mapping under restricted communication Yang et al. (2024); Wang et al. (2025); Asgharivaskasi et al. (2025). Conformal methods provide risk control for adaptive data collection, motion prediction, and reachability Angelopoulos et al. (2024); Xu et al. (2024); Dixit et al. (2023); Muthali et al. (2023), and support help requests, decision calibration, teleoperation,

and control feasibility Ren et al. (2023); Lekeufack et al. (2024); Zhao et al. (2024); Stamouli et al. (2024). These studies motivate our use of held-out calibration, although C-VOIR does not claim a distribution-free conformal guarantee. Existing work does not jointly address quality-aware local task inference, decision-relevant physical probing, and cost-aware team stopping. C-VOIR combines these components while retaining baseline actions and no-op probes when the available evidence does not support execution.

3 Problem Formulation

Task and local beliefs. A fixed robot team $\mathcal{V} = \{1, \dots, n\}$ must identify and execute a task whose deployment reward has an unknown linear representation $\theta \in \mathbb{R}^d$. Robot i observes only local feedback and maintains the posterior

$$\Pi_{i,t}(\theta) = p(\theta \mid \mathcal{I}_{i,t}) \equiv (\mu_{i,t}, \Sigma_{i,t}).$$

Embodiment, viewpoint, sensor noise, and received messages may therefore give robots different beliefs about the same task. At interaction step s, robot i observes

$$y_{i,s} = \phi_{i,s}^\top \theta + \eta_{i,s}, \qquad \mathbb{E}[\eta_{i,s} \mid \phi_{i,s}] = 0, \qquad \mathrm{Var}(\eta_{i,s} \mid \phi_{i,s}) = \sigma_{i,s}^2. \tag{1}$$

For a report r with design matrix X_r, feedback vector y_r, and declared noise level σ_r, we use the whitened statistics

$$\widetilde{X}_r = \frac{X_r}{\sigma_r}, \qquad \widetilde{y}_r = \frac{y_r}{\sigma_r}, \qquad (G_r, h_r, y_{2,r}, n_r) = \left(\widetilde{X}_r^\top \widetilde{X}_r,\ \widetilde{X}_r^\top \widetilde{y}_r,\ \widetilde{y}_r^\top \widetilde{y}_r,\ |y_r|\right). \tag{2}$$

Thus, heterogeneous sensor noise is converted into report-specific precision before quality weighting. The remaining report-message, validity, and deduplication rules are given in Appendix B.

Actions, probes, and feasibility. Robot i has an embodiment-specific action library $\mathcal{A}_i$ induced by candidate BFM policies. Action $a \in \mathcal{A}_i$ has feature $x_{i,a} \in \mathbb{R}^d$ and value $V_i(a; \theta) = \theta^\top x_{i,a}$. A map $L_i : \mathcal{A}_i \to \mathcal{L}$ associates native actions with team-level decision labels, allowing robots to agree on a label while executing different actions. Appendix A.3 gives the full action and label definitions. A shared task vector does not guarantee a common optimal label, because different action libraries may prefer different labels. The team-level utility and compatibility assumption are defined in Appendix A.3. Robot i also has a probe library $\mathcal{P}_i$. Probe p has expected information matrix $H_{i,p} \succeq 0$ and physical cost $\kappa_{i,p} \geq 0$, which may include time, energy, disturbance, and recovery. The probe model is fixed independently of the hidden task and is detailed in Appendix A.4. Each robot has an always-admissible baseline action a_i^0 and no-op probe p_i^0. All selected pairs must satisfy the declared local feasibility test $\mathsf{Safe}_i(a, p \mid a_i^0, \epsilon_i) = 1$, which represents baseline-relative constraints rather than a guarantee of collision-free execution. Moreover, the team authorizes at most one optional probe:

$$\sum_{i \in \mathcal{V}} \mathbf{1}\left[p_{i,t} \neq p_i^0\right] \leq 1. \tag{3}$$

Appendix A.5 defines the feasibility and cost conditions.

Neighbor-only objective. Robots communicate over a fixed, connected, undirected graph $\mathcal{G}_{\mathrm{comm}} = (\mathcal{V}, \mathcal{E}_{\mathrm{comm}})$. Robot i uses only its local history and state together with neighbor messages: $\mathcal{I}_{i,t} = \mathcal{H}_{i,t}^{\mathrm{loc}} \cup \mathcal{S}_{i,t}^{\mathrm{loc}} \cup \mathcal{M}_{i,t}^{\mathrm{nbr}}$. No central process maintains the posterior, assigns probes, or determines when probing stops; Appendix A.6 states the communication assumptions.

At each epoch, the team must select a common label $\ell_t \in \mathcal{L}$, native actions satisfying $L_i(a_{i,t}) = \ell_t$, and either one optional probe or the no-op probes. We address how to fuse reports of different quality, select a decision-relevant probe whose benefit covers its cost, and stop when no worthwhile probe remains. The formal objectives are given in Appendix A.7, and the corresponding C-VOIR components follow in Section 4.

4 C-VOIR

C-VOIR combines quality-aware evidence fusion, decision-relevant probe selection, and cost-aware stopping. Each robot maintains a local posterior, evaluates its native actions and probes, and ex-

changes reports, bids, and decision opinions with its neighbors. The protocol authorizes at most one optional probe per epoch; baseline actions and no-op probes remain available when validation fails.

4.1 Quality-aware information fusion

Using the whitened sufficient statistics defined in the problem formulation, robot i measures the normalized prediction residual of report r and assigns it an age-dependent quality weight:

$$\begin{aligned} z_{i,r} &= \sqrt{\frac{[y_{2,r} - 2\mu_i^\top h_r + \mu_i^\top G_r \mu_i]_+}{\max\{n_r + \mathrm{tr}(G_r \Sigma_i), \nu\}}}, \\ w_{i,r} &= \max\left\{ w_{\min}, e^{-\,\mathrm{age}_r\,/\tau} \min\left(1, \frac{q_{\mathrm{rep}}}{\max\{z_{i,r}, \nu\}}\right)\right\}. \end{aligned} \tag{4}$$

Here, q_{rep} is calibrated on held-out data, τ controls age decay, and $w_{\min}$ is an optional floor. Thus, declared sensor noise controls a report's statistical precision, whereas $w_{i,r}$ discounts stale or unexpected evidence without assigning permanent reliability to its sender.

From the common prior (Λ_0, b_0), robot i recomputes

$$\begin{aligned} \Lambda_i &= \Lambda_0 + \sum_{r \in \mathcal{R}_i} w_{i,r} G_r, \quad & b_i &= b_0 + \sum_{r \in \mathcal{R}_i} w_{i,r} h_r, \\ \Sigma_i &= \Lambda_i^{-1}, & \mu_i &= \Sigma_i b_i. \end{aligned} \tag{5}$$

Reports retain their source identities when forwarded, preventing duplicate evidence from being counted along multiple communication paths. Appendix B gives the report format, processing rules, communication assumptions, and posterior-disagreement bounds.

Fusion property. Robots with the same prior, epoch, weighting configuration, and valid report set compute the same posterior, independently of message arrival order.

4.2 Decision-relevant probe selection

Robot i first filters its action library using its baseline-relative feasibility constraint. For each decision label ℓ, it retains the admissible native action with largest posterior-mean value $\widehat{V}_i(a) = \mu_i^\top x_{i,a}$. Let ℓ_i be the best resulting label and let $\mathcal{D}_i$ contain only unresolved contrasts between ℓ_i and alternative labels. Differences between native actions carrying the same label therefore do not create probe targets.

For feasible probe p, C-VOIR defines

$$\mathrm{NetVOI}_i(p) = \sum_{d \in \mathcal{D}_i} \omega_{i,d} \beta \left[\sqrt{d^\top \Sigma_i d} - \sqrt{d^\top (\Sigma_i^{-1} + H_{i,p})^{-1} d} \right]_+ - \lambda \kappa_{i,p}. \tag{6}$$

The first term measures expected uncertainty reduction along contrasts that can change the selected label; the second charges for physical cost. Each robot publishes at most its highest positive-NetVOI feasible probe. Valid bids are ranked by the total order

$$\mathsf{K}(B) = \big(-\mathrm{NetVOI}(B), \kappa(B), \mathrm{robotID}(B), \mathrm{probeID}(B) \big). \tag{7}$$

The minimum key identifies the candidate winner. Execution additionally requires a strict-majority lease for that exact robot–probe pair; missing confirmation or equivocation cancels the optional probe. Appendix C specifies the active contrasts, bid validity, dissemination, lease construction, and proofs.

Probe property. Robots with the same valid bid set select the same winner, and authenticated non-equivocating robots cannot form strict-majority leases for two different winners in one epoch.

4.3 COST-AWARE DISTRIBUTED STOPPING

Let $v_i = \max_{p:\mathsf{Safe}_i(a_i,p)=1} \mathrm{NetVOI}_i(p)$ be robot i's best admissible probe value, γ_i its conservative decision-label gap, and $\varepsilon_{\mathrm{num}} \geq 0$ a numerical tolerance. The team may stop optional probing only when

$$\begin{aligned}\mathsf{Stop}(t) \iff \mathsf{FullParticipation} \wedge \mathsf{SameEpoch} \wedge \mathsf{Fresh} \wedge \mathsf{NoEquivocation} \\ \wedge \mathsf{SameLabel} \wedge \left[\max_{i\in\mathcal{V}} v_i \leq \varepsilon_{\mathrm{num}}\right].\end{aligned} \tag{8}$$

Thus, stopping requires both statistical evidence that no admissible probe has sufficient net value and protocol agreement on the current state and label. If equation 8 holds and $\min_i \gamma_i > 0$, robots execute native actions for the common label. If the gap is non-positive, they stop optional probing but retain baseline actions and no-op probes without claiming a winner. Missing, stale, conflicting, or equivocating opinions fail closed. Appendix D gives the opinion format, validation order, outputs, and proof.

Stopping property. Under full active participation, an unsupported task action cannot be authorized by stale, conflicting, or incomplete opinions.

Algorithm 1: C-VOIR at one decision epoch

Input: reports $\mathcal{R}_i$, posterior prior, libraries $(\mathcal{A}_i, \mathcal{P}_i)$, epoch t
Output: PROBE, EXECUTE, NODECISION, or FAILCLOSED

for all robots $i \in \mathcal{V}$ **in parallel do**
  Validate, deduplicate, prewhiten, and weight $\mathcal{R}_i$
  Update (μ_i, Σ_i) and compute (ℓ_i, γ_i)
  Compute best feasible probe $(p_i^\star, v_i)$
  Publish $(i, p_i^\star, v_i)$ only if $v_i > \varepsilon_{\mathrm{num}}$
end for
Propagate bids and select $B^\star = \arg\min_B \mathsf{K}(B)$
if a positive winner $B^\star$ exists **then**
  return PROBE iff its exact-pair lease is valid; otherwise **return** FAILCLOSED
end if
Exchange opinions $(\ell_i, \gamma_i, v_i, t)$
if any opinion is missing, stale, conflicting, or equivocal **then**
  return FAILCLOSED
else if $\mathsf{Stop}(t)$ and $\min_i \gamma_i > 0$ **then**
  return EXECUTE for the common label
else
  return NODECISION with baseline actions and no-op probes
end if

5 EXPERIMENTS

We evaluate C-VOIR on standard online task-inference domains, a heterogeneous robot benchmark, and four controlled mechanism studies.

5.1 EVALUATION PROTOCOL

Both main experiments compare Pooled No Probe, Information Gain, C-VOIR, OpTI-UCB, Thompson Sampling, and Random. HeteroArm-Mission additionally includes Local No Probe, a true-task Oracle, and adapted RAGE Fiez et al. (2019) and DecBE-LUCB Amani et al. (2023). Local No Probe uses only local evidence, whereas Pooled No Probe combines available team evidence without probing. RAGE and DecBE-LUCB are adapted because their original formulations do not include our report weighting, physical probe costs, and lease checks.

Within each experiment, methods share the checkpoint, evaluation schedule, and feedback draws. The original OpTI references retain their continuous task-embedding interface, whereas the finite-library methods share candidate actions and probes. We report mean and sample standard deviation

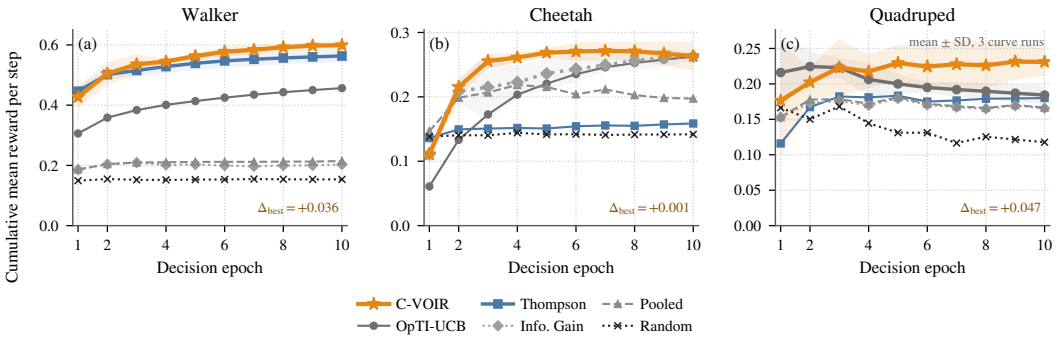

Figure 3: ExORL/DMC running-average return over ten decision epochs. Rewards are first averaged over the four hidden tasks within each trial. Curves show the three-trial mean; bands show sample standard deviation for C-VOIR and the strongest baseline. Δ_{best} is their epoch-10 difference. Exact epoch-10 results are reported in Table 1.

Table 1: Epoch-10 running-average return on ExORL/DMC. Entries are mean ± sample standard deviation over the same three evaluation trials used in Figure 3. Bold denotes the largest reported mean in each domain.

Method	Walker↑	Cheetah↑	Quadruped↑
Random	0.1536 ± 0.0046	0.1418 ± 0.0023	0.1177 ± 0.0145
Pooled No Probe	0.2146 ± 0.0042	0.1973 ± 0.0193	0.1666 ± 0.0284
Information Gain	0.2032 ± 0.0049	0.2628 ± 0.0066	0.1657 ± 0.0290
OpTI-UCB	0.4565 ± 0.0217	0.2624 ± 0.0322	0.1842 ± 0.0132
Thompson Sampling	0.5632 ± 0.0212	0.1587 ± 0.0261	0.1800 ± 0.0170
C-VOIR (ours)	$\mathbf{0.5991 \pm 0.0144}$	$\mathbf{0.2635 \pm 0.0228}$	$\mathbf{0.2311 \pm 0.0196}$

over three independent evaluation trials with training fixed; these trials measure evaluation variability rather than training-seed uncertainty. Full settings and secondary metrics are in Appendix E.

5.2 Experiment A

Setup. ExORL/DMC denotes the exploratory-data evaluation setting built on the DeepMind Control Suite (DMC), a standardized MuJoCo benchmark for continuous control. We use its Walker, Cheetah, and Quadruped locomotion domains, with four hidden reward tasks and one restored FB checkpoint per domain. The embodiment and action interface are fixed within each domain, so Experiment A evaluates online task inference and policy selection rather than embodiment heterogeneity. Each method runs for ten decision epochs per task; an epoch contains 1,000 simulator steps and an optional 20-step probe. C-VOIR and the finite-library baselines use four matched episodes connected in a ring. Returns are divided by 1,000 and are comparable within, but not across, domains.

Results. Figure 3 and Table 1 show that C-VOIR has the largest reported epoch-10 running-average return in all three domains. On Walker, it improves over the strongest baseline, Thompson Sampling, by 0.0359. On Quadruped, it improves over OpTI-UCB by 0.0469. The Cheetah margin over Information Gain is only 0.0007. Overall, the results show that C-VOIR remains competitive with established online task-inference methods while introducing the evidence-fusion and coordination mechanisms needed for heterogeneous robot teams.

5.3 Experiment B

Setup. HeteroArm-Mission contains 600 frozen MuJoCo instances split into 300 training, 150 calibration, and 150 evaluation instances. Each pairs a two-joint arm with a three-joint arm and includes embodiment-specific action libraries, observation-noise standard deviations of 0.08 and 0.16, and physical probes. The benchmark balances precision placement, fast transfer, fragile handling,

Table 2: HeteroArm-Mission results over three trials (mean ± sample standard deviation). Bold marks the best non-Oracle task return, team ϵ-success, and net return. Dashes indicate no physical probe.

Method	$R_{\text{task}} \uparrow$	ϵ-succ.$\uparrow$	Probe cost$\downarrow$	$R_{\text{net}} \uparrow$
Oracle	1.0000 ± 0.0000	1.0000 ± 0.0000	—	1.0000 ± 0.0000
Random	0.4833 ± 0.0100	0.0311 ± 0.0077	—	0.4833 ± 0.0100
Local No Probe	0.8602 ± 0.0024	0.5645 ± 0.0039	—	0.8602 ± 0.0024
Pooled No Probe	0.8701 ± 0.0022	0.6178 ± 0.0102	—	0.8701 ± 0.0022
Information Gain	0.8804 ± 0.0038	0.6467 ± 0.0067	3.1627 ± 0.0665	0.8277 ± 0.0028
OpTI-UCB	0.8712 ± 0.0023	0.6689 ± 0.0038	—	0.8712 ± 0.0023
Thompson	0.8562 ± 0.0020	0.6467 ± 0.0067	—	0.8562 ± 0.0020
DecBE-LUCB	0.8769 ± 0.0065	$\mathbf{0.6978 \pm 0.0154}$	1.3130 ± 0.0190	0.8550 ± 0.0067
RAGE	0.8805 ± 0.0033	0.6511 ± 0.0038	3.2073 ± 0.0296	0.8271 ± 0.0030
C-VOIR(ours)	$\mathbf{0.8810 \pm 0.0034}$	0.6911 ± 0.0077	$\mathbf{0.3597 \pm 0.0405}$	$\mathbf{0.8750 \pm 0.0040}$

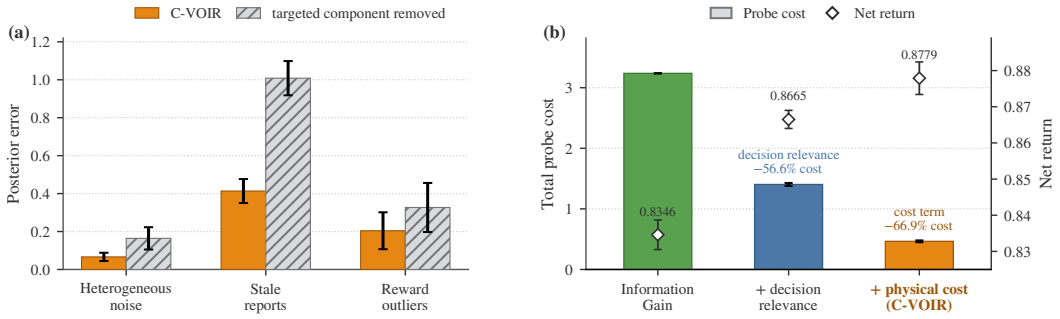

Figure 4: Statistical mechanisms. (a) Posterior error with each quality correction present or removed under its target stress condition; bars show mean ± sample standard deviation over 200 paired seeds. (b) Probe cost (bars) and net return (diamonds) when adding decision relevance and physical cost to the probe objective.

and obstacle handover. Each robot has six task actions, four probes, and a no-op probe; probe cost combines time, energy, terminal-state disturbance, and recovery.

We train HeteroArm-BFM only on the training split and select its checkpoint and C-VOIR hyperparameters on the calibration split. Hidden rewards, task-family labels, and task returns are not model inputs or training targets. Architecture, optimization, data generation, and checkpoint details are provided in Appendix F.

We report normalized task return, team ϵ-success at $\epsilon = 0.1$, physical probe cost, and

$$R^B_{\text{net}} = R^B_{\text{task}} - 0.10\, C_{\text{probe}}/6. \tag{9}$$

The separate local-budget metric in Appendix E.3.2 is not a collision or hardware-failure rate.

Results. Table 2 and Figure 1(c) show that C-VOIR combines the highest non-Oracle task return (0.8810 ± 0.0034) and net return (0.8750 ± 0.0040) with substantially less probing. Its mean probe cost is 88.6% below Information Gain, 88.8% below adapted RAGE, and 72.6% below adapted DecBE-LUCB. Adapted DecBE-LUCB has slightly higher team ϵ-success (0.6978 ± 0.0154 versus 0.6911 ± 0.0077), so C-VOIR improves the performance–cost trade-off rather than every metric.

5.4 Mechanism and Communication Analysis

We isolate C-VOIR's statistical components, stopping predicate, and communication rule in controlled settings whose absolute values are not directly comparable with the main experiments. Protocols and complete tables are in Appendix G.

Removing noise prewhitening, age decay, or residual consistency increases the corresponding posterior error from 0.0663 to 0.1640, from 0.4135 to 1.0084, and from 0.2041 to 0.3268, respectively

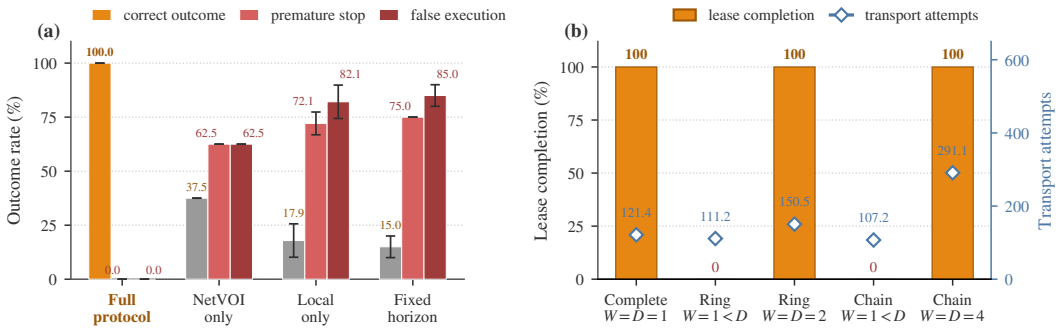

Figure 5: Protocol mechanisms. (a) Outcome rates for four stopping rules over 4,000 constructed cases; error bars aggregate 500 random seeds. (b) Lease completion and transport attempts over 200 paired seeds as the dissemination window varies relative to graph diameter.

(Figure 4(a)). Without age decay, action accuracy also falls from 1.00 to 0.00. These corrections are stress-specific: Appendix G.1 reports that uniform weighting has lower error on clean stationary data without changing the selected action.

Decision-relevant VOI reduces probe cost from 3.2375 to 1.4033; adding the physical-cost term reduces it further to 0.4645 and yields the highest net return, 0.8779 ± 0.0045 (Figure 4(b)). The cost-free variant has slightly higher raw task return, confirming that the cost term changes the operating point rather than improving every metric.

The complete stopping predicate is correct in all constructed states and fails closed in every invalid state. The NetVOI-only rule is correct in 37.5% of cases and prematurely stops or falsely executes in 62.5%. This sweep tests protocol implementation, not physical collision avoidance or unconditional termination.

One communication round suffices on the complete graph, but ring and chain leases fail when the window W is smaller than graph diameter D. Setting $W = D$ restores 100% lease completion and zero posterior disagreement, with mean transport attempts of 150.5 on the ring and 291.1 on the chain. Appendix G.4 reports full results; packet loss and permanent partitions are not simulated here.

Scope and limitations. Experiment A uses standard domains but one shared embodiment, while Experiment B tests heterogeneous actions, observations, and probe costs. HeteroArm-BFM and the adapted RAGE and DecBE-LUCB baselines are not direct executions of their original implementations. Three evaluation trials provide a limited estimate of variability, so we report sample standard deviations and avoid claims of dominance where intervals overlap. The current guarantees assume fixed team membership, a connected communication graph, authenticated identities, declared observation noise, and a bounded decision window.

6 Conclusion

We introduced C-VOIR for online task inference in heterogeneous robot teams. It fuses reports using declared noise, age, and prediction consistency, and selects physical probes by decision-relevant value after accounting for their cost. Conditional analysis shows agreement of posteriors and winning bids, while strict-majority leases prevent conflicting probe authorization under the stated assumptions. Across three evaluation trials, C-VOIR achieves the largest reported epoch-10 running-average return on Walker, Cheetah, and Quadruped. On HeteroArm-Mission, it achieves the highest non-Oracle task return 0.8810 and net return 0.8750, with 88.6% lower probe cost than Information Gain and 72.6% lower cost than adapted DecBE-LUCB. Controlled analyses isolate the effects of report weighting, decision-relevant probing, stopping validation, and communication topology. The evidence is limited to simulation and three evaluation trials; physical deployment requires platform-specific safety and communication validation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of English text. We have not used generative AI tools for scientific idea generation, experimental design, data collection, or result interpretation. We did not use generative AI to produce experimental results, code for core algorithms, or figures.

We reviewed all AI-assisted text edits for factual correctness, technical accuracy, and consistency with the rest of the paper. All scientific claims, methodological decisions, experimental results, and figures were developed and verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses simulated robots and involves no human subjects or personal data. Its safety mechanisms are not a guarantee of safe physical deployment; real-world use requires hardware validation, application-specific risk assessment, secure communication, and human oversight. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The main paper specifies the method, metrics, and evaluation protocol. The appendices provide the full assumptions, proofs, algorithms, dataset splits, hyperparameters, and aggregation rules. The supplementary material provides the code, environment definitions, experiment configurations, and result records used for the tables and figures.

REFERENCES

Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the behavior space of an RL agent. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 566–586, 2025.

Sanae Amani, Tor Lattimore, András György, and Lin Yang. Distributed contextual linear bandits with minimax optimal communication cost. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 691–717, 2023.

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

Arash Asgharivaskasi, Fritz Girke, and Nikolay Atanasov. Riemannian optimization for active mapping with robot teams. *IEEE Transactions on Robotics*, 41:1077–1097, 2025. doi: 10.1109/TRO.2025.3526295.

Xiaoyi Cai, Brent Schlotfeldt, Kasra Khosoussi, Nikolay Atanasov, George J. Pappas, and Jonathan P. How. Energy-aware, collision-free information gathering for heterogeneous robot teams. *IEEE Transactions on Robotics*, 39(4):2585–2602, 2023. doi: 10.1109/TRO.2023.3257512.

Edoardo Cetin, Ahmed Touati, and Yann Ollivier. Finer behavioral foundation models via autoregressive features and advantage weighting. *Reinforcement Learning Journal*, 6:1647–1680, 2025.

Yu-Zhen Janice Chen, Lin Yang, Xuchuang Wang, Xutong Liu, Mohammad Hajiesmaili, John C. S. Lui, and Don Towsley. On-demand communication for asynchronous multi-agent bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3903–3930, 2023.

Anushri Dixit, Lars Lindemann, Skylar X. Wei, Matthew Cleaveland, George J. Pappas, and Joel W. Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pp. 300–314, 2023.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kevin Frans, Seohong Park, Pieter Abbeel, and Sergey Levine. Unsupervised zero-shot reinforcement learning via functional reward encodings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13927–13942, 2024.

Manohari Goarin and Giuseppe Loianno. Graph neural network for decentralized multi-robot goal assignment. *IEEE Robotics and Automation Letters*, 9(5):4051–4058, 2024. doi: 10.1109/LRA.2024.3371254.

Xiao Hu, Jianxiong Li, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. Query-policy misalignment in preference-based reinforcement learning. In *International Conference on Learning Representations*, 2024.

Tyler Ingebrand, Amy Zhang, and Ufuk Topcu. Zero-shot reinforcement learning via function encoders. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21007–21019, 2024.

Pranaya Jajoo, Harshit Sikchi, Siddhant Agarwal, Amy Zhang, Scott Niekum, and Martha White. Regularized latent dynamics prediction is a strong baseline for behavioral foundation models. In *International Conference on Learning Representations*, 2026.

Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 11668–11675, 2024. doi: 10.1109/ICRA57147.2024.10610041.

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. Information directed reward learning for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3850–3862, 2021.

Aritra Mitra, Hamed Hassani, and George J. Pappas. Linear stochastic bandits over a bit-constrained channel. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pp. 1387–1399, 2023.

Anish Muthali, Haotian Shen, Sampada Deglurkar, Michael H. Lim, Rebecca Roelofs, Aleksandra Faust, and Claire J. Tomlin. Multi-agent reachability calibration with conformal prediction. In *Proceedings of the 62nd IEEE Conference on Decision and Control*, pp. 6596–6603, 2023. doi: 10.1109/CDC49753.2023.10383723.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with Hilbert representations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39737–39761, 2024.

Matteo Pirotta, Andrea Tirinzoni, Ahmed Touati, Alessandro Lazaric, and Yann Ollivier. Fast imitation via behavior foundation models. In *International Conference on Learning Representations*, 2024.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121882–121927, 2024. doi: 10.52202/079017-3874.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023.

Hugo Richard, Etienne Boursier, and Vianney Perchet. Constant or logarithmic regret in asynchronous multiplayer bandits with limited communication. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 388–396, 2024.

Thomas Rupf, Marco Bagatella, Marin Vlastelica, and Andreas Krause. Optimistic task inference for behavior foundation models. In *International Conference on Learning Representations*, 2026.

Sudeep Salgia and Qing Zhao. Distributed linear bandits under communication constraints. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29845–29875, 2023.

Harshit Sikchi, Andrea Tirinzoni, Ahmed Touati, Yingchen Xu, Anssi Kanervisto, Scott Niekum, Amy Zhang, Alessandro Lazaric, and Matteo Pirotta. Fast adaptation with behavioral foundation models. *Reinforcement Learning Journal*, 6:1048–1074, 2025.

Igor Spasojevic, Xu Liu, Alejandro Ribeiro, George J. Pappas, and Vijay Kumar. Active collaborative localization in heterogeneous robot teams. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.112.

Charis Stamouli, Lars Lindemann, and George J. Pappas. Recursively feasible shrinking-horizon MPC in dynamic environments with conformal prediction guarantees. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1330–1342, 2024.

Jingbo Sun, Songjun Tu, Qichao Zhang, Haoran Li, Xin Liu, Yaran Chen, Ke Chen, and Dongbin Zhao. Unsupervised zero-shot reinforcement learning via dual-value forward-backward representation. In *International Conference on Learning Representations*, 2025.

Zhuoli Tian, Yuyang Zhang, Jinsheng Wei, and Meng Guo. iHERO: Interactive human-oriented exploration and supervision under scarce communication. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.115.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. In *International Conference on Learning Representations*, 2025.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *International Conference on Learning Representations*, 2023.

Aaron D. Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2147–2156, 2023.

Jun Wang, Guocheng He, and Yiannis Kantaros. Probabilistically correct language-based multi-robot planning using conformal prediction. *IEEE Robotics and Automation Letters*, 10(1):160–167, 2025. doi: 10.1109/LRA.2024.3504233.

Xuchuang Wang, Lin Yang, Yu-Zhen Janice Chen, Xutong Liu, Mohammad Hajiesmaili, Don Towsley, and John C. S. Lui. Exploration for free: How does reward heterogeneity improve regret in cooperative multi-agent bandits? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2192–2202, 2023.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yupeng Yang, Yiwei Lyu, Yanze Zhang, Sha Yi, and Wenhao Luo. Decentralized multi-robot line-of-sight connectivity maintenance under uncertainty. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.005.

Michelle D. Zhao, Reid Simmons, Henny Admoni, and Andrea Bajcsy. Conformalized teleoperation: Confidently mapping human inputs to high-dimensional robot actions. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.008.

A Full Problem Setup and Notation

This appendix gives the definitions omitted from Section 3. It defines the shared task representation, local observation model, posterior state, action and probe libraries, physical costs, local feasibility condition, and neighbor-only information boundary. Message formats, bid propagation, probe leases, and stopping rules are specified in later appendices.

A.1 Team and shared task representation

The fixed robot team is

$$\mathcal{V} = \{1, \dots, n\}. \tag{10}$$

The team must identify and execute one task whose reward is unknown at deployment. Its shared linear representation is

$$\theta \in \mathbb{R}^d. \tag{11}$$

Every robot infers the same θ from its local observations and received messages. Robot identities and team membership remain fixed during the decision process studied in this paper. Dynamic membership and automatic task decomposition are outside the current scope.

A.2 Local observations and posterior state

At interaction step s, robot $i \in \mathcal{V}$ observes a feature vector $\phi_{i,s} \in \mathbb{R}^d$ and scalar task feedback

$$y_{i,s} = \phi_{i,s}^\top \theta + \eta_{i,s}, \tag{12}$$

where $\eta_{i,s}$ is conditionally zero-mean observation noise with declared standard deviation $\sigma_{i,s} > 0$. The noise level may depend on the robot, sensor, and observation type. Each registered noise profile fixes these values before evaluation. The report-weighting method later accounts for observations that depart from this declared noise model.

Robot i stores its local observation history

$$\mathcal{H}_{i,t}^{\mathrm{loc}} = \{(\phi_{i,s}, y_{i,s}) : 0 \le s \le t\}. \tag{13}$$

Its posterior is

$$\Pi_{i,t}(\theta) = p(\theta \mid \mathcal{I}_{i,t}). \tag{14}$$

Under the linear-Gaussian representation, the posterior is described by a precision matrix $\Lambda_{i,t} \succ 0$ and a natural parameter $b_{i,t}$:

$$\Sigma_{i,t} = \Lambda_{i,t}^{-1}, \qquad \mu_{i,t} = \Lambda_{i,t}^{-1} b_{i,t}. \tag{15}$$

All robots begin with the same positive-definite prior (Λ_0, b_0). Their later posterior states can differ because they may have received different observations or neighbor messages.

A report r contains a batch with design matrix X_r, feedback vector y_r, and declared noise standard deviation $\sigma_r > 0$. Observations with different noise profiles are placed in separate reports. We first whiten the batch as

$$\widetilde{X}_r = \frac{X_r}{\sigma_r}, \qquad \widetilde{y}_r = \frac{y_r}{\sigma_r}. \tag{16}$$

The report stores the sufficient statistics

$$G_r = \widetilde{X}_r^\top \widetilde{X}_r = \sigma_r^{-2} X_r^\top X_r, \qquad h_r = \widetilde{X}_r^\top \widetilde{y}_r = \sigma_r^{-2} X_r^\top y_r,$$
$$y_{2,r} = \widetilde{y}_r^\top \widetilde{y}_r = \sigma_r^{-2} y_r^\top y_r, \qquad n_r = |y_r|. \tag{17}$$

The whitened observation noise has unit variance. These statistics are enough for the weighted linear posterior update used in Section 4; in particular, a larger declared σ_r produces a smaller precision contribution for the same raw observations. This appendix defines their statistical meaning. Their message identity, validity, forwarding, and replay rules belong to the decentralized method rather than the problem definition.

A.3 Action libraries and team decision labels

Robot i has a finite library of robot-native task actions induced by its candidate BFM policies:

$$\mathcal{A}_i = \{a_{i,1}, \ldots, a_{i,K_i}\}. \tag{18}$$

Each action $a \in \mathcal{A}_i$ has a task-independent feature vector $x_{i,a} \in \mathbb{R}^d$ and value

$$V_i(a;\theta) = \theta^\top x_{i,a}. \tag{19}$$

The physical actions of two robots need not match. We therefore define a decision-label map

$$L_i : \mathcal{A}_i \to \mathcal{L}, \qquad \ell_{i,a} = L_i(a), \tag{20}$$

where $\mathcal{L}$ is the set of team-level policy choices. Robots agree on a decision label, while each robot may execute a different embodiment-specific action associated with that label. This mapping separates agreement on what the team should do from the native control used by each robot to carry it out. The label map does not by itself imply a common optimum. For an active decision epoch, let $\mathcal{A}_i^{\mathrm{adm}}$ denote the locally admissible action set and define

$$\mathcal{L}^{\mathrm{com}} = \bigcap_{i \in \mathcal{V}} L_i(\mathcal{A}_i^{\mathrm{adm}}). \tag{21}$$

We define the team utility of a common label as

$$U(\ell;\theta) = \sum_{i \in \mathcal{V}} \alpha_i \max_{a \in \mathcal{A}_i^{\mathrm{adm}}:\, L_i(a) = \ell} V_i(a;\theta), \qquad \alpha_i \geq 0, \qquad \sum_{i \in \mathcal{V}} \alpha_i = 1. \tag{22}$$

The coefficients α_i are fixed team-level utility weights.

C-VOIR assumes that the active libraries contain an acceptable common label with a positive team-utility gap. Let

$$\ell^\star \in \arg\max_{\ell \in \mathcal{L}^{\mathrm{com}}} U(\ell;\theta), \qquad \gamma^U = U(\ell^\star;\theta) - \max_{\ell \in \mathcal{L}^{\mathrm{com}} \setminus \{\ell^\star\}} U(\ell;\theta) > 0. \tag{23}$$

If $\mathcal{L}^{\mathrm{com}}$ is empty or the gap is not positive, the protocol does not claim a task winner and retains the baseline actions and no-op probes. This is an explicit compatibility assumption on the active action libraries, not a consequence of sharing the same task vector.

A.4 Probe information and physical cost

Robot i has a finite probe library

$$\mathcal{P}_i = \{p_{i,1}, \ldots, p_{i,M_i}, p_i^0\}, \tag{24}$$

where p_i^0 is the no-op probe. A probe is a short action whose purpose is to collect information about θ. Examples include observing from another viewpoint, checking whether a route is traversable, or testing whether an object can be moved.

Probe p has a predicted information matrix

$$H_{i,p} \in \mathbb{R}^{d\times d}, \qquad H_{i,p} \succeq 0. \tag{25}$$

The matrix represents the expected increase in posterior precision if the probe produces its planned observation, in the same noise-whitened units as G_τ. If $Z_{i,p}$ is the raw predicted probe design matrix and $\sigma_{i,p}$ is its declared observation-noise standard deviation, then $H_{i,p} = \sigma_{i,p}^{-2} Z_{i,p}^\top Z_{i,p}$. The probe library and its information matrices are constructed without access to the hidden θ or the unobserved reward outcome. The later selection rule may use the current posterior to choose among these fixed candidates.

The physical-cost vector of probe p is

$$\mathbf{c}_{i,p} = \left(c_{i,p}^{\text{time}}, c_{i,p}^{\text{energy}}, c_{i,p}^{\text{disturbance}}, c_{i,p}^{\text{recovery}}\right)^\top \in \mathbb{R}_+^4. \tag{26}$$

The terms measure elapsed time, energy use, change to the current task state, and the effort needed to resume normal task execution. Fixed nonnegative conversion weights $\boldsymbol{\rho}_i$ map these components to a scalar cost certificate:

$$\kappa_{i,p} = \boldsymbol{\rho}_i^\top \mathbf{c}_{i,p} \geq 0. \tag{27}$$

The conversion weights are fixed before evaluation. The realized normalized cost is assumed not to exceed $\kappa_{i,p}$ on the stated probe-cost certificate event. For the no-op probe,

$$H_{i,p_i^0} = 0, \qquad \kappa_{i,p_i^0} = 0. \tag{28}$$

A.5 Baseline action and local feasibility

At decision epoch t, robot i has a declared baseline action

$$a_{i,t}^0 \in \mathcal{A}_i. \tag{29}$$

The baseline is an action that can continue under the robot's current information and safety rules. It may be a normal operational action or a declared holding action when task execution is not yet supported.

Let

$$C_{i,t}(a \mid a_{i,t}^0, \Pi_{i,t}) \geq 0 \tag{30}$$

denote a certificate for the possible task loss caused by selecting a instead of the baseline. Section 4 instantiates this term with a posterior confidence bound. Let $\mathcal{F}_{i,t}^{\text{phys}}$ contain action–probe pairs allowed by the robot's platform-level constraints. Then

$$\mathsf{Safe}_i(a, p \mid a_{i,t}^0, \epsilon_i) = 1 \tag{31}$$

means that

$$(a,p) \in \mathcal{F}_{i,t}^{\text{phys}} \quad \text{and} \quad C_{i,t}(a \mid a_{i,t}^0, \Pi_{i,t}) + \kappa_{i,p} \leq \epsilon_i. \tag{32}$$

The budget ϵ_i is local to robot i. A low average cost across the team cannot make an infeasible pair admissible for one robot. Here, Safe_i denotes the declared platform constraints and the baseline-relative budget test in Eq. equation 32; it does not by itself certify collision-free physical execution. We require

$$C_{i,t}(a^0_{i,t} \mid a^0_{i,t}, \Pi_{i,t}) = 0, \qquad \mathsf{Safe}_i(a^0_{i,t}, p^0_i \mid a^0_{i,t}, \epsilon_i) = 1. \tag{33}$$

Thus, the baseline action and no-op probe remain feasible from the first decision epoch. In addition, the team-level mutual-exclusion constraint in Eq. equation 3 allows at most one non-no-op probe in each decision epoch. Local feasibility therefore determines which robot–probe pairs may be considered, while team coordination determines whether one of them is authorized.

A.6 Communication and local information

Robots communicate over

$$\mathcal{G}_{\mathrm{comm}} = (\mathcal{V}, \mathcal{E}_{\mathrm{comm}}), \tag{34}$$

which is fixed, connected, and undirected during the decision window studied in the main guarantees. The neighbors of robot i are

$$\mathcal{N}_i = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}_{\mathrm{comm}}\}. \tag{35}$$

Connectivity allows information to reach every robot after enough communication rounds, but it does not give any robot immediate access to global information.

The information available to robot i is

$$\mathcal{I}_{i,t} = \mathcal{H}^{\mathrm{loc}}_{i,t} \cup \mathcal{S}^{\mathrm{loc}}_{i,t} \cup \mathcal{M}^{\mathrm{nbr}}_{i,t}, \tag{36}$$

where $\mathcal{M}^{\mathrm{nbr}}_{i,t}$ contains messages received through $\mathcal{N}_i$ and $\mathcal{S}^{\mathrm{loc}}_{i,t}$ contains the robot's current platform state. Every action, probe, bid, and stopping opinion produced by robot i must depend only on $\mathcal{I}_{i,t}$ and fixed public parameters.

We do not assume a central node that observes all local histories, maintains a global posterior, or directly assigns probes. A centralized posterior may be used as an evaluation reference, but it is not available to the deployed method.

Our paper studies fixed membership and a static connected graph. Its exact agreement guarantees use synchronous communication windows. Bounded packet loss is handled only under a stated delivery bound. Arbitrary asynchronous convergence, permanent network partitions, Byzantine report fabrication, dynamic membership, and time-varying communication graphs are outside the current scope.

A.7 Decision objectives

The team must solve three coupled problems.

First, robots must use local evidence from different sources without treating all observations as equally informative. The fusion rule should account for report quality while avoiding false confidence caused by communication.

Second, the robots must decide whether an optional probe reduces uncertainty that can still affect the current behavior choice. A probe should be selected only when it satisfies the local feasibility test and its decision benefit is large enough to cover its physical cost.

Third, the team must stop probing when no feasible probe has positive net decision value. Because no central node observes every local state, this statistical condition must be combined with a distributed agreement condition. Section 4 gives the three corresponding C-VOIR rules.

B Report Fusion and Consistent Report Processing

This appendix specifies the report representation, acceptance rules, common posterior reprocessing procedure, and the guarantees used in Section 4.1.

B.1 INCREMENTAL REPORT REPRESENTATION

For a local observation batch with feature matrix X_r, feedback vector y_r, and registered noise standard deviation σ_r, robot $s(r)$ uses the whitening rule in Eq. equation 16 and computes

$$G_r = \widetilde{X}_r^\top \widetilde{X}_r, \qquad h_r = \widetilde{X}_r^\top \widetilde{y}_r, \qquad y_{2,r} = \widetilde{y}_r^\top \widetilde{y}_r, \qquad n_r = |y_r|. \tag{37}$$

The first two statistics update the linear posterior. The last two allow a receiver to measure standardized prediction error without receiving the raw observations. The immutable noise profile identifies the registered σ_r used to construct these statistics. An incremental report has the form

$$\begin{aligned} r = (\,&\mathrm{reportID}, \mathrm{taskID}, \mathrm{senderID}, \mathrm{sequenceNo}, \mathrm{createdAt}, \mathrm{expiresAt}, \\ &G_r, h_r, y_{2,r}, n_r, \mathrm{noiseProfile}, \mathrm{probeID}, \mathrm{probeCost}, \mathrm{signature}). \end{aligned} \tag{38}$$

The source identifier is bound to the immutable origin fields:

$$\mathrm{reportID} = (\mathrm{taskID}, \mathrm{senderID}, \mathrm{sequenceNo}). \tag{39}$$

A forwarding robot may change routing metadata, such as the hop count, but it does not change the source identifier or the evidence fields. Thus, a forwarded report remains the same observation rather than becoming new evidence.

B.2 ACCEPTANCE, DEDUPLICATION, AND EXPIRATION

At epoch t, robot i accepts report r only if all of the following hold:

1. its task instance matches the receiver;
2. the dimensions of (G_r, h_r) match the local posterior;
3. $\mathrm{createdAt}(r) \le t$;
4. either no expiration is declared or $t < \mathrm{expiresAt}(r)$;
5. the source-bound report ID is valid; and
6. the same report ID has not already been accepted with different content.

The receiver stores a digest of the immutable fields. A repeated ID with the same digest is a duplicate and does not update the posterior. A repeated ID with different content is an ID conflict and is rejected. Expired reports stay in the processed-ID archive, so receiving an old copy later cannot make the same evidence active again.

The acceptance outcomes used by the implementation are shown in Table 3.

Outcome	Meaning
`accepted`	The report is valid and is added to the archive.
`duplicate`	The same ID and content were already stored.
`report_id_conflict`	The same ID names different content.
`wrong_task_instance`	The task instance does not match.
`wrong_dimension`	The sufficient statistics have the wrong size.
`future_report`	The creation epoch is later than the local epoch.
`expired`	The report has reached or passed its expiration epoch.

Table 3: Report acceptance outcomes. Rejected reports never change the posterior.

B.3 CONSISTENT REPORT PROCESSING

Report quality depends on the receiver's current posterior. Applying reports immediately in network arrival order would therefore make the final result depend on communication timing. At each decision epoch, a robot instead takes all valid archived reports and processes them according to the shared key

$$(\text{createdAt}, \text{senderID}, \text{sequenceNo}, \text{reportID}). \tag{40}$$

Starting from the common prior (Λ_0, b_0), the reprocessing pass recomputes each report weight using Eq. equation 4 and applies the information-form update in Eq. equation 5 exactly once for each report under this shared rule. If the weight of a later report depends on earlier updates, that weight is recomputed during every pass. The posterior is not formed by applying a new weight to an already accumulated archive.

Lemma 1 (arrival-order invariance). Suppose two robots use the same prior, decision epoch, weighting configuration, valid report set, shared processing rule, and deterministic numerical rules. They produce the same ordered weight sequence and the same $(\Lambda, b, \mu, \Sigma)$, regardless of the order in which the network delivered the reports.

Proof. The common immutable report set and sorting key produce the same sequence $(r_1, \ldots, r_m)$. Before r_1, both processing states equal the same prior. If the states before r_k agree, the deterministic weighting rule receives the same report, posterior, age, and configuration, and therefore returns the same weight. The posterior update then produces the same next state. Induction over the sequence proves the claim. □

B.4 Propagation assumptions and report-set agreement

For the formal communication result, fix one decision epoch and a finite, connected, undirected graph $\mathcal{G}_{\text{comm}} = (\mathcal{V}, \mathcal{E})$. The graph remains fixed during the decision window. In one synchronous round, a sender transmits its round-start report set to every neighbor. Reports are immutable, an accepted ID is retained during the window, and a missing report is treated as having zero weight. Let $d_{\mathcal{G}}(i, j)$ be shortest-path distance and let D be the graph diameter.

Let $\mathcal{R}_i^{(k)}$ denote robot i's valid report-ID set after k communication rounds.

Theorem 1 (diameter-time report agreement). Under reliable synchronous delivery,

$$\mathcal{R}_i^{(k)} = \bigcup_{j:\, d_{\mathcal{G}}(i,j) \le k} \mathcal{R}_j^{(0)}. \tag{41}$$

Consequently, after D rounds every robot holds the same union $\mathcal{R}^\star = \bigcup_j \mathcal{R}_j^{(0)}$. Each source report contributes at most once despite duplicate paths or cycles.

Proof. The identity is immediate for $k = 0$. In round $k + 1$, robot i keeps its set and unions the round-start sets of its neighbors. Applying the induction hypothesis gives exactly the reports whose origins are at most $k + 1$ hops away. All pairwise distances are at most D, and insertion by immutable ID is idempotent. □

The reliable-delivery assumption can be weakened to bounded blackout. Suppose that every block of B consecutive rounds contains at least one successful transmission on each directed edge for every report already held by the sender, and that reports are retransmitted every round.

Theorem 2 (bounded-blackout delivery). If a report is valid at epoch t and

$$\text{expiresAt}(r) \ge t + DB, \tag{42}$$

then it reaches every robot no later than the end of round $DB - 1$. Hence all robots hold the same valid report union within DB rounds.

Proof. Along a shortest path of length $\ell \le D$, bounded blackout delivers the report across each successive edge within at most B rounds. Retransmission keeps it available after every successful hop, so delivery takes at most $\ell B \le DB$ rounds. The TTL condition keeps the report valid throughout this window. □

This theorem does not provide liveness under permanent packet loss, an insufficient TTL, graph disconnection, or an unbounded asynchronous delay.

B.5 Centralized special case and finite-communication difference

Define the centralized reference as a process that collects the valid union $\mathcal{R}^\star$ and applies exactly the same prior, weighting rule, report-processing procedure, local action and probe computations, and deterministic winner key as the robots.

Corollary 1 (centralized special case). After report-set agreement, every local posterior equals the centralized posterior. After a further diameter-length bid dissemination window, every robot also selects the same probe as the reference. If no positive bid exists, all systems select the no-op probe.

Before report-set agreement, different report ledgers can produce different posteriors. Extend each ledger by setting $w_{i,r} = 0$ when robot i does not hold r, and define

$$\Delta\Lambda_{ij} = \sum_r (w_{i,r} - w_{j,r}) G_r, \qquad \Delta b_{ij} = \sum_r (w_{i,r} - w_{j,r}) h_r. \tag{43}$$

Let $\lambda_i = \lambda_{\min}(\Lambda_i) > 0$ and $\lambda_j = \lambda_{\min}(\Lambda_j) > 0$.

Theorem 3 (posterior difference certificate). The posterior differences satisfy

$$\|\mu_i - \mu_j\|_2 \le B_{ij}^{\mu} := \frac{\|\Delta b_{ij}\|_2 + \|\Delta\Lambda_{ij}\|_2 \|\mu_j\|_2}{\lambda_i}, \tag{44}$$

$$|x^\top(\mu_i - \mu_j)| \le \|x\|_2 B_{ij}^{\mu}, \tag{45}$$

$$\|\Sigma_i - \Sigma_j\|_2 \le \frac{\|\Delta\Lambda_{ij}\|_2}{\lambda_i \lambda_j}. \tag{46}$$

Proof. Because $\Lambda_i \mu_i = b_i$ and $\Lambda_j \mu_j = b_j$,

$$\Lambda_i(\mu_i - \mu_j) = \Delta b_{ij} - \Delta\Lambda_{ij}\mu_j. \tag{47}$$

Multiplying by Λ_i^{-1} and applying the triangle inequality gives equation 44; Cauchy–Schwarz gives equation 45. The resolvent identity $\Sigma_i - \Sigma_j = \Lambda_i^{-1}(\Lambda_j - \Lambda_i)\Lambda_j^{-1}$ gives equation 46. □

The certificate separates two claims that should not be confused: consistent report processing removes arrival-order dependence once report sets agree, while equation 44–equation 46 quantify the remaining difference when communication has not yet made those sets equal.

C Probe Selection, Bidding, and Lease Uniqueness

This appendix gives the local decision certificates, active contrast set, complete bidding procedure, and lease proof used in Section 4.2.

C.1 Baseline-relative action certificate

Let a_i^0 be robot i's baseline action and let a be a candidate action. For the feature contrast $c_{i,a} = x_{i,a_i^0} - x_{i,a}$, define the conservative relative-loss certificate

$$\delta_i(a \mid a_i^0) = \left[c_{i,a}^\top \mu_i + \beta \sqrt{c_{i,a}^\top \Sigma_i c_{i,a}} + e_i \|c_{i,a}\|_2 \right]_+, \tag{48}$$

where e_i is a bound on value-model error. Robot i admits action a only if $\delta_i(a \mid a_i^0) \le \epsilon_i$. Among admitted actions, it chooses the largest posterior-mean value, with a fixed action-ID tie rule. Since $\delta_i(a_i^0 \mid a_i^0) = 0$, the baseline action is always admitted.

Define the admissible action and label sets as

$$\mathcal{A}_i^{\mathrm{adm}} = \{a \in \mathcal{A}_i : \delta_i(a \mid a_i^0) \le \epsilon_i\}, \qquad \mathcal{L}_i^{\mathrm{adm}} = \{L_i(a) : a \in \mathcal{A}_i^{\mathrm{adm}}\}. \tag{49}$$

A probe p with physical-cost certificate $\kappa_{i,p}$ is feasible together with action a_i only if

$$\delta_i(a_i \mid a_i^0) + \kappa_{i,p} \le \epsilon_i. \tag{50}$$

This is a per-robot hard constraint. One robot cannot compensate for another robot exceeding its local budget.

C.2 Active contrasts and local probe value

Let $L_i : \mathcal{A}_i \to \mathcal{L}$ map a robot-native action to its team-level decision label. For each available label $\ell \in \mathcal{L}_i^{\mathrm{adm}}$, define the representative native action

$$a_i^\star(\ell) = \arg \max_{a \in \mathcal{A}_i^{\mathrm{adm}}:\, L_i(a)=\ell} \widehat{V}_i(a), \tag{51}$$

using a fixed action-ID tie rule. The selected label and action are

$$\ell_i \in \arg \max_{\ell \in \mathcal{L}_i^{\mathrm{adm}}} \widehat{V}_i(a_i^\star(\ell)), \qquad a_i = a_i^\star(\ell_i), \tag{52}$$

with a fixed label tie rule. For every alternative label $\ell \in \mathcal{L}_i^{\mathrm{adm}} \setminus \{\ell_i\}$, define

$$d_{i,\ell} = x_{i,a_i^\star(\ell)} - x_{i,a_i}, \qquad m_{i,\ell} = \widehat{V}_i(a_i) - \widehat{V}_i(a_i^\star(\ell)), \tag{53}$$

and

$$u_{i,\ell} = \beta \sqrt{d_{i,\ell}^\top \Sigma_i d_{i,\ell}} + q_{\mathrm{cal}} \|d_{i,\ell}\|_2. \tag{54}$$

Here, β is the posterior-confidence multiplier and q_{cal} is a value-model error bound fixed on held-out calibration data.

The active set contains only unresolved alternative labels:

$$\mathcal{D}_i = \{d_{i,\ell} : \ell \in \mathcal{L}_i^{\mathrm{adm}} \setminus \{\ell_i\},\ m_{i,\ell} \le u_{i,\ell}\}. \tag{55}$$

The corresponding weight is

$$\omega_{i,\ell} = \frac{[u_{i,\ell} - m_{i,\ell}]_+}{\max(u_{i,\ell}, \nu)}. \tag{56}$$

All quantities are recomputed from the current posterior at each decision epoch. Native actions assigned to the same label are therefore not counted as different decision alternatives.

For $H_{i,p} \succeq 0$, the covariance predicted after probe p is

$$\Sigma_i'(p) = (\Sigma_i^{-1} + H_{i,p})^{-1}. \tag{57}$$

The resulting decision-relevant uncertainty reduction is

$$R_i(p) = \sum_{d_{i,\ell} \in \mathcal{D}_i} \omega_{i,\ell} \beta \left[\sqrt{d_{i,\ell}^\top \Sigma_i d_{i,\ell}} - \sqrt{d_{i,\ell}^\top \Sigma_i'(p) d_{i,\ell}} \right]_+ \tag{58}$$

and $\mathrm{NetVOI}_i(p) = R_i(p) - \lambda \kappa_{i,p}$ as in equation 6. If the active set is empty, then $R_i(p) = 0$ for every probe. If a probe is infeasible, its value is set to $-\infty$ and it cannot be published as a bid.

Because $H_{i,p} \succeq 0$, $\Sigma_i'(p) \preceq \Sigma_i$. Hence each summand in equation 58 is nonnegative. This does not imply that the net value is positive: the probe is useful only when its weighted decision-width reduction is larger than its certified cost.

The current probe value is intentionally robot-local: it measures the reduction in unresolved decision-label uncertainty for the robot evaluating the probe. Consequently, a probe with little local value for robot i may still be useful for robot j but may not be published under the one-bid rule. A team-aggregated probe value would require an additional cross-robot utility model and is outside the present protocol.

C.3 One-bid rule and bid validity

Let $\mathcal{P}_i^{\mathrm{adm}}$ be the probes satisfying equation 50. Robot i chooses

$$p_i^\star \in \arg\max_{p \in \mathcal{P}_i^{\mathrm{adm}}} \mathrm{NetVOI}_i(p), \tag{59}$$

with lower cost and then probe ID as local tie breakers. It publishes a bid only when $\mathrm{NetVOI}_i(p_i^\star) > \varepsilon_{\mathrm{num}}$. Therefore, each robot sends at most one positive bid in a decision epoch. This is enough for global selection because the maximum over the union of local probe libraries equals the maximum of the local maxima.

The published bid is

$$B_i = (\mathrm{taskID}, t, \mathrm{robotID}, \mathrm{probeID}, \mathrm{NetVOI}, \mathrm{costCertificate}, \mathrm{expiresAt}). \tag{60}$$

A receiver considers the bid valid only if its task and decision epoch match the active decision; its expiration is later than the current epoch; its net VOI is larger than $\varepsilon_{\mathrm{num}}$; and its numeric fields are finite with nonnegative cost. Bids from a different decision epoch cannot compete with current bids.

All valid bids use the total order in equation 7. The order is fixed before execution and is independent of network arrival time.

C.4 Bid dissemination

Let $\mathcal{B}_i^{(k)}$ be robot i's valid bid view after k synchronous rounds. Under the same static connected graph assumption used in Appendix B, bid views satisfy

$$\mathcal{B}_i^{(k)} = \bigcup_{j:\, d_{\mathcal{G}}(i,j) \le k} \mathcal{B}_j^{(0)}. \tag{61}$$

Theorem 4 (exact winner after dissemination). After k rounds, robot i selects the best bid originating in its k-hop neighborhood. After D rounds, every robot selects the globally best positive bid. If there is no positive bid, every robot retains the no-op probe.

Proof. Equation equation 61 follows by the same induction as the report propagation result. Applying the same total winner order to the same set gives the same minimum key. After D rounds, every bid is present at every robot. Since every robot publishes its local maximum, the best published bid is also the best probe in the union of the local feasible libraries. □

Before D rounds, this rule is exact only over the visible neighborhood. With no bound or spatial regularity assumption on bid values, topology alone cannot give a finite additive approximation to an unseen bid: an arbitrarily valuable probe can be placed just outside the current communication radius.

C.5 Acknowledgement and lease state machine

The bid winner is an exact pair $(\mathrm{robotID}^\star, \mathrm{probeID}^\star)$. Each participant that accepts this winner publishes

$$A_j = (\text{taskID}, t, \text{senderID}_j, \text{winnerRobotID}, \text{winnerProbeID}). \tag{62}$$

Acknowledgements are counted by distinct authenticated sender IDs. A sender that acknowledges two different winner pairs for the same task and decision epoch has equivocated; the lease attempt is then rejected. If the accepted acknowledgement set contains at least q_{lease} matching senders, the system creates

$$\mathsf{Lease} = (\text{taskID}, t, \text{robotID}^\star, \text{probeID}^\star, \text{confirmedBy}). \tag{63}$$

Only the robot named in Lease may execute the probe named in Lease. A bid by itself is not permission to act. The protocol therefore has the following states:

1. *No winner*: no valid positive bid exists; keep the no-op probe.
2. *Winner pending*: a winner exists, but acknowledgements are still incomplete; do not execute an optional probe.
3. *Lease confirmed*: at least q_{lease} distinct participants confirm the same exact pair; only the named robot executes it.
4. *Lease cancelled*: confirmation expires, a required participant is missing, or equivocation is observed; keep the baseline action and no-op probe.

For the uniqueness result below, participant identities are authenticated and each honest participant acknowledges at most one winner pair per task and decision epoch.

Theorem 5 (unique strict-majority lease). Let $m = |\mathcal{V}|$. If $q_{\text{lease}} > m/2$, at most one winner pair can obtain a valid lease in one decision epoch. The default full-participation rule $q_{\text{lease}} = m$ is a special case.

Proof. Suppose two different winner pairs obtained acknowledgement sets Q_1 and Q_2, each of size at least q_{lease}. Since $|Q_1| + |Q_2| > m$, the two sets intersect. A participant in the intersection would have acknowledged two different pairs in the same epoch, contradicting non-equivocation. □

If $q_{\text{lease}} \le m/2$, two disjoint participant sets can acknowledge different winners. A smaller quorum therefore needs an additional consensus, locking, or membership protocol; changing only the numeric threshold does not preserve the unique-lease theorem. Likewise, hidden equivocation by unauthenticated or Byzantine participants lies outside the stated model.

D TEAM STOPPING PROTOCOL

This appendix specifies the opinion message, validation order, outputs, and fail-closed guarantee used in Section 4.3.

D.1 DECISION OPINIONS

For one task and decision epoch t, robot i publishes

$$O_i = (\text{taskID}, t, \text{robotID}, \ell_i, a_i, \gamma_i, v_i, s_i), \tag{64}$$

where ℓ_i is the team-level decision label, a_i is the robot-native action for that label, γ_i is its conservative label gap, v_i is the net VOI of its best feasible probe, and s_i is the state epoch. Different robots may report different native action IDs while agreeing on the same decision label.

For the selected label ℓ_i, the conservative label gap is

$$\gamma_i = \min_{\ell \neq \ell_i} \left\{ \Delta_{i,\ell}^\top \mu_i - \beta \sqrt{\Delta_{i,\ell}^\top \Sigma_i \Delta_{i,\ell}} - e_i \|\Delta_{i,\ell}\|_2 \right\}, \qquad \Delta_{i,\ell} = x_{i,a_i} - x_{i,a_i^\star(\ell)}. \tag{65}$$

The minimum is over alternative decision labels, not over native actions with the same label. If no alternative label is available, we set $\gamma_i = +\infty$. On the simultaneous contrast-confidence event, $\gamma_i > 0$ means that the selected decision label is separated from every alternative label represented in the current local action library. A non-positive gap prevents the protocol from claiming reliable label identification, even if all feasible probes have non-positive net value.

An opinion is atomic: v_i equals the maximum net value over all locally admissible probes, and a bid exists exactly when $v_i > \varepsilon_{\text{num}}$. The local posterior, action library, and state epoch used to form the opinion do not change between publication and evaluation of that stopping round.

D.2 Validation order

Let $\mathcal{Q}(t)$ be the accepted opinions for the team. The default rule requires one opinion from every robot. The checks are applied in the following order.

1. Context and sender uniqueness. Ignore opinions from another task, decision epoch, or non-participant. If the same sender supplies two different opinions for the current context, return `conflicting_opinions`.

2. Participation. Require at least q_{stop} distinct participants. The theorem below uses full participation, $q_{\text{stop}} = |\mathcal{V}|$. If the threshold is not reached, return `insufficient_confirmations`.

3. State freshness. For every accepted opinion, require

$$0 \le t_{\text{now}} - s_i \le A_{\max}. \tag{66}$$

If $s_i > t_{\text{now}}$, return `future_state`. If $t_{\text{now}} - s_i > A_{\max}$, return `stale_state`.

4. Decision-label agreement. Require $\ell_i = \ell_j$ for all accepted opinions. If more than one label is present, return `decision_disagreement`. Agreement is checked on the shared decision label rather than the robot-native action ID.

5. No positive local probe. Require

$$\max_{i \in \mathcal{V}} v_i \le \varepsilon_{\text{num}}. \tag{67}$$

If this check fails, return `positive_probe_voi` and continue the decision process. This condition is the cost-aware statistical stopping rule: no checked robot has an admissible probe whose decision benefit covers its physical cost.

6. Action certificate. After all earlier checks pass, compute $\gamma_{\min} = \min_i \gamma_i$. If $\gamma_{\min} > 0$, stop probing and authorize the native actions associated with the common label. Otherwise, stop optional probing but retain the baseline action and no-op probe.

D.3 Outputs and reason codes

The stopping evaluator returns

$$(\text{shouldStop}, \text{reason}, \text{decisionLabel}, \text{actionMap}, \text{executeActions}, \gamma_{\min}). \tag{68}$$

Its outcomes are summarized in Table 4.

The first six outcomes fail closed: they issue no stopping certificate and authorize no optional probe or unsupported task action. The system waits for a new decision epoch or for fresher, complete, and consistent information.

Reason	Stop probing	Execute task actions
`conflicting_opinions`	No	No
`insufficient_confirmations`	No	No
`future_state`	No	No
`stale_state`	No	No
`decision_disagreement`	No	No
`positive_probe_voi`	No	No
`certified_consensus`	Yes	Yes
`uncertain_consensus_noop`	Yes	No

Table 4: Team stopping outcomes. "No" in the last column means that the protocol retains the baseline/no-op fallback and does not authorize an unsupported team task decision.

D.4 Fail-closed stopping guarantee

Assume a frozen decision epoch, full team participation, authenticated and non-equivocating senders, atomic opinions, and fresh state epochs.

Theorem 6 (fail-closed distributed stopping). The stopping rule has the following properties:

1. If any active participant has $v_i > \varepsilon_{\mathrm{num}}$, the team does not stop.
2. If the team stops, all active participants report one common decision label and none reports a positive feasible probe.
3. The team executes the reported native actions only when $\min_i \gamma_i > 0$; otherwise it stops probing and retains the baseline action and no-op probe.
4. Any conflicting, stale, future, or incomplete opinion set cannot authorize a task action.

On the simultaneous local contrast-confidence event, $\min_i \gamma_i > 0$ also implies that every robot's reported native action has higher true value than each challenger in its declared local library.

Proof. Full participation places every current v_i in the checked set. The rule explicitly rejects stopping if any value exceeds $\varepsilon_{\mathrm{num}}$, which proves the first claim. The preceding context, freshness, and label checks prove the second and fourth claims. After these checks, action execution is authorized exactly when the minimum conservative gap is positive; otherwise the result is `uncertain_consensus_noop`, proving the third claim.

For the final statement, each expression inside the minimum in equation 65 is a lower confidence bound on the true value difference between the selected action and one challenger. A positive minimum makes every such true contrast positive. □

Quorum limitation. With fewer than all team opinions, an omitted robot may still have a positive bid or a different decision label. The theorem therefore does not apply below full participation. Failure-tolerant stopping would require a separate membership or failure-detection protocol, which is outside the present scope.

E Main-Experiment Details

This appendix gives the evaluation protocol, dataset-level parameters, and secondary metrics omitted from the main paper. Unless stated otherwise, reported values are the mean and sample standard deviation over three evaluation trials.

E.1 Aggregation and reporting conventions

For Experiment A, let $r_{i,t,e}$ denote the return per simulator step, averaged over the four matched robot episodes, in evaluation trial i, task t, and decision epoch e. We first average the four hidden

tasks within each trial and epoch, and then compute the running average over decision epochs:

$$\bar{r}_{i,e} = \frac{1}{4}\sum_{t=1}^{4} r_{i,t,e}, \qquad c_{i,e} = \frac{1}{e}\sum_{s=1}^{e} \bar{r}_{i,s}. \tag{69}$$

Figure 3 reports the mean and sample standard deviation of $c_{i,e}$ over the three evaluation trials at each decision epoch e. Table 1 in the main paper reports the corresponding epoch-10 values $c_{i,10}$. We divide the raw DMC return by the 1,000-step episode horizon. These values are average rewards per simulator step, not success percentages or Oracle-normalized scores.

A trial changes the evaluation random stream while keeping the trained checkpoint, hidden-task list, method parameters, and evaluation budget fixed. For Experiment A, trial $i \in \{0, 1, 2\}$ uses the base evaluation seed plus $100{,}000i$. Thus, the reported sample standard deviation measures evaluation variation rather than variation across independently trained models.

For Experiment B, each trial evaluates the same 150 frozen mission instances. We first average each metric over the instances in one trial, then report the mean and sample standard deviation of the three trial-level values. Team ϵ-success requires both robots in one mission instance to finish with normalized regret at most $\epsilon = 0.1$. Probe cost is accumulated over the six decision rounds. Budget exceedance records whether the realized loss relative to the baseline action, including probe cost, exceeds the declared local relative-loss budget. It does not denote collision or hardware failure.

E.2 Experiment A: ExORL/DMC

E.2.1 Protocol and parameters

We evaluate Walker, Cheetah, and Quadruped with four hidden reward tasks per domain. Each domain uses one FB checkpoint restored at 2,000,000 updates. The evaluation uses base seed 20260915, three trials, ten decision epochs, four robots, and a ring communication graph. Each epoch has 1,000 task-action steps. A physical probe uses a 20-step prefix. Reports and bids propagate for two communication rounds.

We choose C-VOIR parameters at the domain level rather than for individual tasks. Walker and Cheetah use the original configuration. Quadruped uses the more exploratory configuration selected during development. All four tasks in a domain share the same values in Table 5.

Table 5: Dataset-level C-VOIR parameters for Experiment A.

Parameter	Walker	Cheetah	Quadruped
`noise_std`	0.01	0.01	1.0
`beta`	1.0	1.0	1.0
`online_exploration_scale`	1.0	1.0	2.0
`online_candidate_count`	8	8	32
`online_update_steps`	20	20	20
Local budget ϵ_i (`safety_epsilon`)	2.0	2.0	2.0

OpTI-UCB uses $\beta = 1$, 128 candidates, and prior precision 1. Thompson Sampling uses sampling scale 0.01. These parameters are specific to their respective algorithms. What is shared across methods is the restored checkpoint, task instance, feedback realization, decision horizon, and interaction budget.

E.3 Experiment B: HeteroArm-Mission

E.3.1 Dataset, model, and parameters

HeteroArm-Mission contains 600 frozen MuJoCo mission instances split into 300 training, 150 calibration, and 150 evaluation instances. Each instance contains a two-joint arm and a three-joint arm with different action libraries and observation-noise standard deviations of 0.08 and 0.16. We balance four task families within each split: precision placement, fast transfer, fragile handling, and obstacle handover. Each robot has six task actions, four physical probes, and a no-op probe. Probe

cost combines time, energy, and terminal-state disturbance, while each probe trajectory includes a recovery segment.

Because the source data contain candidate-policy trajectories rather than the transition buffer required by the original FB training procedure, we train HeteroArm-BFM, a task-independent successor-feature model with a shared scene encoder, learned candidate embeddings, and separate action-feature and probe-trajectory heads. The model is trained on the 300 training instances and selected by loss on the 150 calibration instances; evaluation instances are not used for checkpoint selection. Hidden task weights, task-family labels, and task returns are not model inputs or training targets. On the evaluation split, its action-feature relative error is 0.4162, its probe-information relative error is 0.6842, and its zero-shot action accuracy is 0.5967. Calibration selects $\beta = 1.0$ and a VOI cost weight of 0.06 before method evaluation. Full data generation, model architecture, training, checkpoint selection, and metric definitions are given in Appendix F.

Table 6: Frozen parameters for Experiment B.

Parameter	Value
Evaluation trials	3
Decision rounds	6
Posterior β	1.0
VOI cost weight	0.06
Probe-return penalty	0.10
Local relative-loss budget ϵ_i	1.0
Success tolerance ϵ	0.10
Communication rounds	1
Prior precision	fitted on training split
BFM training seed	0
Scene input dimension	18
Hidden dimension	128
Embedding dimension	16
Action-feature dimension	12
Probe horizon	16
Maximum training epochs	400
Selected epoch index	140
Early-stopping patience	60
Batch size	256
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-5}
Probe-loss weight	1.0
Gradient-norm limit	5.0

E.3.2 Local-budget exceedance metric

Table 7 reports the three-trial budget-exceedance rate reported separately from the main HeteroArm table. The Oracle is included as an upper bound, not as a deployable inference method.

The budget metric evaluates the declared loss certificate. It should not be interpreted as a physical accident rate. C-VOIR's value of 0.0318 ± 0.0017 is lower than Thompson Sampling and DecBE-LUCB, but higher than the no-probe baselines and OpTI-UCB.

F HeteroArm-Mission and HeteroArm-BFM Details

F.1 Dataset construction

HeteroArm-Mission is constructed from frozen HeteroArm action and probe trajectories. Dataset construction does not regenerate or alter those physical trajectories. It assigns hidden reward vectors and noisy scalar feedback to the frozen candidates. The 600 mission instances are split into 300

Table 7: Local-budget exceedance rate on HeteroArm-Mission. Entries are mean ± sample standard deviation over three evaluation trials.

Method	Budget exceedance↓
Oracle	0.0000 ± 0.0000
Local No Probe	0.0026 ± 0.0006
Pooled No Probe	0.0126 ± 0.0017
Information Gain	0.0278 ± 0.0011
C-VOIR	0.0318 ± 0.0017
OpTI-UCB	0.0200 ± 0.0011
Thompson Sampling	0.0463 ± 0.0006
DecBE-LUCB, adapted	0.0504 ± 0.0006
RAGE, adapted	0.0326 ± 0.0017
Random	0.5022 ± 0.0011

training, 150 calibration, and 150 evaluation instances. Task families are balanced separately within every split.

Let $\bar{\theta}_f \in \mathbb{R}^{12}$ be the fixed template for task family f. For mission instance g assigned to family $f(g)$, the hidden task vector is

$$\widetilde{\theta}_g = \bar{\theta}_{f(g)} + \xi_g, \qquad \xi_g \sim \mathcal{N}(0, 0.035^2 I), \qquad \theta_g = \frac{\widetilde{\theta}_g}{\|\widetilde{\theta}_g\|_2}. \tag{70}$$

Here, g indexes an independent dataset instance, not an algorithmic subtask; the two robots in instance g share the same θ_g. The random draws use dataset seed 20260916. The four family templates, ordered by the twelve task features, are

$$\bar{\theta}_{\text{precision}} = (0, 0, -.15, -.15, 0, .15, -.75, -.65, .65, .35, .75, .10), \tag{71}$$

$$\bar{\theta}_{\text{transfer}} = (0, 0, 0, 0, 0, 0, -.15, .90, 1.00, .45, .20, 0), \tag{72}$$

$$\bar{\theta}_{\text{fragile}} = (0, 0, -.10, -.10, 0, .20, -1.00, -1.00, .45, .45, .90, .15), \tag{73}$$

$$\bar{\theta}_{\text{handover}} = (.15, .20, 0, 0, 0, .10, -.30, -.15, .45, .75, .35, 1.20). \tag{74}$$

The feature order is end-effector x, end-effector y, their squared terms, first-joint cosine, posture, control energy, end-effector speed, payload x- and y-displacement, payload proximity, and obstacle clearance. For candidate action a, the noise-free task value and observed scalar feedback are

$$V_g(a) = x_{g,a}^\top \theta_g, \qquad y_{g,a} = V_g(a) + \epsilon_{g,a}, \qquad \epsilon_{g,a} \sim \mathcal{N}(0, \sigma^2_{e(a)}), \tag{75}$$

where $e(a) \in \{0, 1\}$ identifies the embodiment and $(\sigma_0, \sigma_1) = (0.08, 0.16)$. Probe feedback uses the same embodiment-specific noise levels. All methods receive the same frozen scene instances, task schedule, and feedback-generation rule within an evaluation trial. Each probe trajectory contains an excursion followed by recovery; its reported cost combines duration, energy, and terminal-state disturbance.

F.2 HeteroArm-BFM architecture

HeteroArm-BFM predicts task-independent quantities for the fixed candidate libraries. Its public input for an action candidate consists of four scene variables, a 14-dimensional policy descriptor, candidate type, embodiment ID, and behavior ID. The policy descriptor contains the policy target, two policy offsets, six-dimensional policy identity, two-dimensional embodiment identity, and two controller parameters. Probe candidates use the same 18-dimensional input layout with a zero policy-descriptor block. Hidden task weights, task family, expected task values, observed task values, and task returns are excluded from the input.

After componentwise normalization using training instances only, an 18-dimensional candidate input s is processed by a two-layer scene encoder:

$$h_s = \text{MLP}_{\text{scene}}(s), \qquad 18 \to 128 \to 128, \tag{76}$$

with SiLU activations. Candidate type, embodiment, and behavior each have a learned 16-dimensional embedding. Their concatenation with h_s is processed by a shared behavior encoder,

$$h = \mathrm{MLP}_{\mathrm{shared}}\left([h_s; e_{\mathrm{type}}; e_{\mathrm{emb}}; e_{\mathrm{behavior}}]\right), \qquad 176 \to 128 \to 128, \tag{77}$$

again with SiLU activations. A linear action head predicts a 12-dimensional successor-feature vector $\widehat{x}_{g,a}$. A separate linear probe head predicts a length-16 trajectory of 12-dimensional features, $\widehat{Z}_{g,i,p} \in \mathbb{R}^{16\times 12}$. The predicted raw probe Gram matrix is

$$\overline{H}_{g,i,p} = \sum_{t=0}^{15} 0.98^t (\widehat{z}_t - \widehat{z}_0)(\widehat{z}_t - \widehat{z}_0)^\top. \tag{78}$$

The action and probe outputs are denormalized before evaluation. No-op probe outputs are set to zero.

F.3 Training and checkpoint selection

All inputs and targets are normalized componentwise using means and standard deviations computed from the 300 training instances. Let $\bar{x}$ denote a normalized action-feature target and let $\bar{Z}$ denote a normalized flattened probe-trajectory target. The training objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}_a|} \sum_{(g,a)\in\mathcal{B}_a} \|\widehat{\bar{x}}_{g,a} - \bar{x}_{g,a}\|_2^2 + \lambda_{\mathrm{probe}} \frac{1}{|\mathcal{B}_p|} \sum_{(g,i,p)\in\mathcal{B}_p} \|\widehat{\bar{Z}}_{g,i,p} - \bar{Z}_{g,i,p}\|_F^2, \tag{79}$$

with $\lambda_{\mathrm{probe}} = 1$. We train with AdamW, learning rate 10^{-3}, weight decay 10^{-5}, action and probe batch size 256, and gradient norm clipping at 5. Training uses seed 0 for at most 400 epochs. After every epoch, we compute the same weighted loss on the 150 calibration instances. We save the checkpoint with the lowest calibration loss, using an improvement threshold of 10^{-7}, and stop after 60 epochs without improvement. The reported model was selected at epoch index 140 with calibration loss 0.47805; training stopped after 221 epochs. The 150 evaluation instances were not used to optimize model parameters or select this checkpoint.

On the evaluation split, we report

$$E_{\mathrm{action}} = \frac{\|\widehat{X} - X\|_F}{\|X\|_F} = 0.4162, \qquad E_{\mathrm{probe}} = \frac{\|\widehat{H} - H\|_F}{\|H\|_F} = 0.6842. \tag{80}$$

Zero-shot action accuracy is the fraction of embodiment-specific candidate sets for which the action maximizing $\widehat{x}_{g,a}^\top \theta_g$ matches the action maximizing $x_{g,a}^\top \theta_g$. It is used only as an evaluation metric; θ_g is not a model input. The evaluation accuracy is 0.5967.

F.4 HeteroArm evaluation metrics

For mission instance g, robot i, and decision round t, let $a_{g,i,t}$ be the selected native action and let $V_{g,i}(a)$ be its noise-free value under the hidden task vector. Define the value range

$$\Delta_{g,i} = \max_{a\in\mathcal{A}_{g,i}} V_{g,i}(a) - \min_{a\in\mathcal{A}_{g,i}} V_{g,i}(a), \tag{81}$$

with a numerical floor of 10^{-12}. The normalized instantaneous regret and return are

$$r_{g,i,t} = \frac{\max_a V_{g,i}(a) - V_{g,i}(a_{g,i,t})}{\Delta_{g,i}}, \qquad \rho_{g,i,t} = \frac{V_{g,i}(a_{g,i,t}) - \min_a V_{g,i}(a)}{\Delta_{g,i}} = 1 - r_{g,i,t}. \tag{82}$$

For $M = 2$ robots and $T = 6$ rounds, task normalized return is

$$R_{\mathrm{task},g} = \frac{1}{MT} \sum_{t=1}^{T} \sum_{i=1}^{M} \rho_{g,i,t}. \tag{83}$$

If C_g is total probe cost and $\alpha = 0.10$, the reported cost-adjusted return and cumulative regret are

$$R_{\mathrm{net},g} = R_{\mathrm{task},g} - \alpha \frac{C_g}{T}, \qquad \mathrm{Regret}_g = \sum_{t=1}^{T} \sum_{i=1}^{M} r_{g,i,t} + \alpha C_g. \tag{84}$$

Let $a_{g,i,T}$ be the final selected action. Its final normalized regret is

$$r_{g,i}^{\text{final}} = \frac{\max_a V_{g,i}(a) - V_{g,i}(a_{g,i,T})}{\Delta_{g,i}}. \tag{85}$$

For tolerance $\epsilon = 0.10$, team ϵ-success is

$$S_g^{\epsilon} = \mathbb{I}\left[\max_{i\in\{1,2\}} r_{g,i}^{\text{final}} \le \epsilon\right]. \tag{86}$$

Reported dataset-level metrics first average these instance quantities over the 150 evaluation instances. The paper then reports the mean and sample standard deviation of the resulting trial-level aggregates over the three evaluation trials.

To measure compatibility of the common-label assumption, let $L_{g,i}$ denote robot i's label map, let $\widehat{\ell}_{g,i,T}$ be its final locally selected label before agreement validation, let $\mathcal{A}_{g,i,T}^{\text{adm}}$ be its final admissible action set, and let $o_{g,T}$ be the team protocol output. Define

$$\begin{aligned} I_g^{\text{common}} &= \mathbb{I}\left[\bigcap_{i=1}^{2} L_{g,i}(\mathcal{A}_{g,i,T}^{\text{adm}}) \neq \varnothing\right], \\ I_g^{\text{conflict}} &= \mathbb{I}\left[I_g^{\text{common}} = 1 \ \wedge\ \widehat{\ell}_{g,1,T} \neq \widehat{\ell}_{g,2,T}\right], \\ I_g^{\text{fallback}} &= \mathbb{I}[o_{g,T} \in \{\text{NoDecision}, \text{FailClosed}\}]. \end{aligned} \tag{87}$$

For $N = 150$ evaluation instances, report

$$\begin{aligned} \widehat{q}_{\text{common}} &= \frac{1}{N}\sum_{g=1}^{N} I_g^{\text{common}}, \qquad \widehat{q}_{\text{fallback}} = \frac{1}{N}\sum_{g=1}^{N} I_g^{\text{fallback}}, \\ \widehat{q}_{\text{conflict}} &= \frac{\sum_{g=1}^{N} I_g^{\text{conflict}}}{\max\left(\sum_{g=1}^{N} I_g^{\text{common}}, 1\right)}. \end{aligned} \tag{88}$$

These are the common-label feasibility, conditional label-conflict, and fallback/no-decision rates. They should be reported alongside task return and team success. They are distinct from team ϵ-success: a team may satisfy the per-robot regret threshold while still having incompatible local label preferences.

G Mechanism and Communication Analysis

No robot reads a global posterior, a global bid table, or a centralized stopping flag. Report fusion, bid ranking, lease formation, and stopping use only local state and neighbor-to-neighbor messages. We isolate the three statistical components of C-VOIR and the effect of communication topology. Variants within each comparison use paired task instances and report draws. These controlled analyses are reported separately from the main table because their absolute values use experiment-specific stress settings.

G.1 Quality-aware report fusion

We study a controlled linear-Gaussian problem with five robots, a four-dimensional task vector, and eight report epochs. All robots reprocess accepted reports using the shared rule in Appendix B. We use 200 paired seeds and separately introduce heterogeneous observation noise, stale reports after a task shift, and reward outliers. Targeted variants remove noise prewhitening, age decay, or residual consistency. Uniform weighting removes age decay and residual weighting but retains noise prewhitening. Posterior error is $\|\widehat{\theta} - \theta\|_2 / \|\theta\|_2$ at the final epoch.

Noise prewhitening lowers posterior error when robot noise levels differ. Age decay prevents stale evidence from dominating after a task shift, and residual consistency reduces error under reward outliers. These corrections are not uniformly best in every benign case: with clean stationary data, uniform weighting has lower posterior error than the full rule (0.0702 ± 0.0267 versus 0.0890 ± 0.0340), while both select the correct action. The weighting rule is therefore useful under the targeted quality failures, not a guarantee of lower estimation error on every distribution.

Table 8: Report-fusion ablation over 200 paired seeds. Entries are mean ± sample standard deviation.

Stress	Fusion rule	Posterior error↓	Action acc.↑
Heterogeneous noise	C-VOIR	0.0663 ± 0.0218	1.000 ± 0.000
	Without prewhitening	0.1640 ± 0.0589	1.000 ± 0.000
	Uniform report weights	0.0543 ± 0.0201	1.000 ± 0.000
Stale task reports	C-VOIR	0.4135 ± 0.0632	1.000 ± 0.000
	Without age decay	1.0084 ± 0.0903	0.000 ± 0.000
	Uniform report weights	0.7043 ± 0.0630	0.540 ± 0.500
Reward outliers	C-VOIR	0.2041 ± 0.0973	0.995 ± 0.071
	Without residual consistency	0.3268 ± 0.1294	0.995 ± 0.071
	Uniform report weights	0.2685 ± 0.0995	0.995 ± 0.071

Table 9: Probe-objective ablation on HeteroArm-Mission. Entries are mean ± sample standard deviation over three paired evaluation trials.

Variant	$R_{\text{task}} \uparrow$	ϵ-succ.↑	Probes↓	Cost↓	$R_{\text{net}} \uparrow$
No probe	0.873 ± 0.001	0.627 ± 0.012	—	—	0.873 ± 0.001
Task-agnostic IG	0.889 ± 0.004	0.658 ± 0.020	5.998 ± 0.004	3.238 ± 0.002	0.835 ± 0.004
Decision VOI	0.890 ± 0.002	0.687 ± 0.007	2.616 ± 0.043	1.403 ± 0.024	0.867 ± 0.003
C-VOIR	0.886 ± 0.005	0.676 ± 0.015	0.884 ± 0.027	0.465 ± 0.015	0.878 ± 0.005

G.2 Decision relevance and physical probe cost

We use the 150 HeteroArm-Mission evaluation instances and the frozen HeteroArm-BFM checkpoint. Within each of three evaluation trials, all variants share mission instances and feedback draws. No Probe uses pooled evidence without optional probing. Task-agnostic Information Gain uses log-determinant information gain. Decision VOI without cost uses only active decision-label contrasts and sets the cost coefficient to zero. Full C-VOIR additionally charges physical probe cost. Because this is a separately paired mechanism study, its absolute values are not combined with Table 2.

Restricting information gain to active decision contrasts reduces probe cost by 56.6% relative to task-agnostic Information Gain while improving team success. Charging physical cost reduces probe cost by a further 66.9% relative to cost-free Decision VOI and gives the highest net return. Cost-free Decision VOI has slightly higher raw task return and team success; the cost term changes the performance–cost trade-off rather than improving every outcome.

G.3 Team stopping checks

We evaluate four rules on eight balanced protocol states with 500 seeds per state. The states represent certified consensus, a positive probe value, an insufficient decision gap, a missing participant, stale and future opinions, a conflicting label, and equivocation. NetVOI Only omits opinion validation, Local Only uses one opinion, and Fixed Horizon stops after six rounds. A correct outcome must match both the stop and action-authorization decision specified by the constructed state. Fail closed is conditioned on the five invalid-opinion states: missing participation, stale or future state, label conflict, and equivocation.

The full predicate executes only for fresh, complete, non-conflicting opinions with a common label. It continues when positive feasible NetVOI remains, stops without an action claim when the label gap is insufficient, and rejects action execution in invalid-opinion states. This controlled state sweep checks the implemented predicate; it does not establish collision avoidance or unconditional termination.

G.4 Communication topology and dissemination window

We use five robots connected by complete, ring, and chain graphs of diameter one, two, and four. We compare one communication round with a dissemination window equal to the graph diameter

Table 10: Stopping-protocol state sweep over 4,000 cases per rule. Rates are percentages; fail closed is conditional on the five invalid-opinion states.

Rule	Correct↑	Premature stop↓	False execution↓	Fail closed↑
Full protocol	100.0	0.0	0.0	100.0
NetVOI Only	37.5	62.5	62.5	0.0
Local Only	17.9	72.1	82.1	0.0
Fixed Horizon	15.0	75.0	85.0	0.0

Table 11: Effect of communication topology over 200 paired seeds. Requested runs count seeds with at least one lease attempt; completed/failed counts all lease attempts. Other entries are mean $\pm$ sample standard deviation.

Graph	D	Win.	Req.	Comp./fail.	Lease↑	Disagr.↓	Transport↓
Complete	1	1	73	$76/0$	1.000	0.000 ± 0.000	121.4 ± 147.0
Ring	2	1	199	$0/294$	0.000	0.370 ± 0.169	111.2 ± 122.0
Ring	2	2	73	$76/0$	1.000	0.000 ± 0.000	150.5 ± 159.2
Chain	4	1	199	$0/355$	0.000	0.504 ± 0.252	107.2 ± 101.8
Chain	4	4	73	$76/0$	1.000	0.000 ± 0.000	291.1 ± 273.5

over 200 paired seeds. Lease completion is the number of completed leases divided by completed plus failed lease attempts; runs with no requested lease do not affect this rate. Transport attempts count report deliveries and received bid copies rather than serialized bytes.

One round is sufficient only on the complete graph. Matching the dissemination window to the graph diameter completes every requested lease and eliminates posterior disagreement on both sparse graphs, but increases transport work. This experiment does not include packet loss, TTL expiration, or permanent network partitions.