

Who to Trust in Multi-Robot Systems: Joint Robot and Object Hypothesis Trust Inference via Neural-Symbolic Cross-Validation

Ruiyang Wang, Hao-Lun Hsu, Haocheng Meng, Songqiao Xie and Miroslav Pajic

Abstract—Multi-robot systems (MRS) offer significant advantages over single-robot platforms by enabling information sharing and coordination for different tasks in complex environments. However, reliable cooperation critically depends on inferring trust over both information sources and their reported hypotheses, particularly under partial observability and potential adversarial interference. Existing approaches frequently depend on pre-established trust relationships or side-channel information, which may not be available or reliable in realistic MRS deployments. This work introduces *NeST-Bayes*, a neural-symbolic framework that jointly infers robot and hypothesis trust using only inter-robot observation consistency, without requiring preassigned robot trust, reputation information, or side-channel identity signals. *NeST-Bayes* infers instantaneous trust evidence scores for robots and object hypotheses by reasoning over ego-centric relational graphs constructed from cross-robot observation consistency. These evidence scores are accumulated over time through Bayesian updates of Beta-distributed trust beliefs to maintain temporally consistent trust estimates. Extensive evaluations show that *NeST-Bayes* achieves the highest mean object precision and robot recall across all benchmark settings, and the highest robot precision and object recall in most settings, compared to trust-based methods. Under aggressive optimized attacks, it achieves 98.8%/93.8% robot precision/recall and 98.3%/95.4% object precision/recall. In an occlusion-rich Webots warehouse, it maintains 92.5%/92.0% robot precision/recall and 97.3%/85.0% object precision/recall. This shows joint improvements in the reliability and coverage of retained information.

I. INTRODUCTION

By enabling coordinated decision-making and information exchange among multiple robots, multi-robot systems (MRS) achieve performance gains over single-robot platforms in perception, exploration, coverage, and planning. Seminal works [1], [2] established the value of cooperation in distributed robot teams, while more recent efforts improved perceptual robustness through multi-view sensing [3] and enhanced search efficiency via information sharing and coordinated planning [4].

While multi-viewpoint aggregation overcomes individual sensing limitations, real-world MRS must operate under partial observability and sensing uncertainty [5], unreliable communication [6], and restricted fields of view [7]. More critically, shared estimation and planning become vulnerable in contested environments and scenarios, where adversarial robots may inject/utilize misleading information; e.g., Sybil attacks distort collective objectives through false identities [8], and Byzantine behaviors corrupt distributed diffusion and state estimation [9], [10]. Any mechanism aggregating inter-robot information should therefore infer trust over both the robots providing reports and the hypotheses they produce.

Existing trust inference approaches in MRS typically rely on assumptions that limit their applicability in realistic settings. Early works [11]–[13] treat trust as a pre-established or

reputation-based quantity, while more adaptive methods [14], [15] infer it from hand-designed models or specifications of nominal behavior. Recent resilient hypothesis-testing methods [16], [17] exploit side-channel information that is often unavailable. Complementary sensor-fusion approaches reject inconsistent measurements without explicitly estimating the reliability of their sources [18], [19]. Bayesian methods such as MATE [20] already use observation consistency to jointly estimate agent and track trust, with each estimate informing the other. However, their mappings from consistency checks to trust evidence rely on hand-designed rules. A critical limitation is the lack of methods to learn how typed consistency and contradiction relations can be combined, to infer robot and object hypothesis trust under potential adversarial reporting.

Consequently, in this work, we introduce *NeST-Bayes*, a Neural Symbolic Trust inference framework that learns relational evidence to jointly estimate robot and object-hypothesis trust from cross-robot observation consistency, with Bayesian temporal evidence accumulation through updates to Beta-distributed trust beliefs. For each robot, we construct an ego-centric heterogeneous graph whose typed edges encode observation agreement and contradiction using geometric and field-of-view (FoV) checks. A triplet encoder initializes node representations from their symbolic relation profiles, and relation-specific graph attention propagates evidence across robot and hypothesis nodes to produce instantaneous trust scores. These scores are accumulated as fractional pseudo-counts in Beta distributions to maintain temporal trust beliefs.

NeST-Bayes requires no preassigned robot trust or side-channel identity signals at inference time. Evaluations under noisy sensing, normal and optimized adversarial reporting, increased adversarial ratios, and occlusion-rich Webots simulations show that *NeST-Bayes* generally achieves the highest precision while also leading in recall among trust-based methods. Component ablation studies show that relational propagation improves recall while maintaining high precision.

In summary, the main contributions of this work include:

- A neural-symbolic trust inference framework that combines symbolic relation profiles with heterogeneous graph attention to jointly infer robot and object-hypothesis evidence, integrating the learned scores with Bayesian temporal evidence accumulation.
- An extensive evaluation against *No Trust*, *Naïve Bayesian*, and *MATE* baselines under sensing noise, normal and optimized adversarial reporting, increased adversarial ratios, and occlusion-rich Webots simulations, together with component ablations demonstrating how symbolic relation initialization and typed relational propagation contribute to precision and recall.

II. RELATED WORK

A. MRS Vulnerabilities

MRS are sensitive to sensing noise and partial observability [21], communication uncertainty [22], [23], and malicious interference [24]. False data injection can destabilize coordination and navigation [25], and adversarial perception attacks inject spurious measurements that disrupt cooperative decision-making [19]. Byzantine behaviors in distributed estimation and learning have also been studied [26], [27], where even a few compromised robots degrade performance by injecting false positives or omitting valid observations. These results motivate principled trust mechanisms that operate under limited information and unreliable sensing, without assuming nominal-mode behavior.

B. Trust in MRS

Trust estimation in MRS remains an open problem. Existing approaches derive trust evidence from explicitly defined behaviors or task specifications [15], [28], or exploit side-channel information that effectively fingerprints the robot identity [16], [17]. More recent methods [20], [29], [30] cross-validate observations across robots, but rely on hand-designed rules for trust updates that, as we show, can be brittle when adversaries are numerous or cross-robot validation is sparse. *NeST-Bayes* instead learns how symbolic consistency and contradiction relations combine through heterogeneous message passing, producing joint robot and object-hypothesis evidence for temporal trust updates.

C. Neural-Symbolic AI

Neural-symbolic methods combine the interpretability of symbolic reasoning with the representational flexibility of neural networks [31], [32], and have proven effective for logical inference [33], probabilistic logic programming [34], and relational reasoning [35]. In robotics, they have succeeded in task planning [36], visual and relational reasoning [37], [38], and manipulation [39], [40]. Neural Logic Machines (NLMs) [41] learn multi-step relational reasoning by composing symbolic predicates over a shared entity domain. Trust inference in MRS instead involves heterogeneous robot and object-hypothesis entities connected by typed observation and consistency relations. Relation-aware graph neural networks with heterogeneous message passing [42]–[44] provide a sparse substrate for this setting. *NeST-Bayes* is conceptually related to NLM-style relational reasoning, but operates over ego-centric heterogeneous graphs rather than using the explicit predicate operators of an NLM.

III. PROBLEM FORMULATION

We consider a multi-robot system $\mathcal{R} = \{r_i\}_{i=1}^M$ operating in a workspace $\mathcal{W} \subset \mathbb{R}^d$ over a finite horizon $t \in \{0, \dots, T\}$. The environment contains a set of dynamic objects $\mathcal{O}$, where each object $o \in \mathcal{O}$ is represented by its position state $\mathbf{p}_o(t) \in \mathcal{W}$ and unique identifier $id(o) \in \mathcal{I}_{\text{gt}}$. Each robot r_i has state

$$\mathbf{s}_i(t) = (\mathbf{p}_i(t), \theta_i(t), R_i^{\text{sen}}, \Phi_i), \quad (1)$$

where $\mathbf{p}_i(t)$ is its position, $\theta_i(t)$ is its heading, and $(R_i^{\text{sen}}, \Phi_i)$ define its range- and angle-limited FoV, namely $A_i(t)$. The visible ground-truth objects are

$$\mathcal{O}_i(t) = \{o \in \mathcal{O} \mid \mathbf{p}_o(t) \in A_i(t)\}. \quad (2)$$

A. Noisy Sensing and Local Information

Each robot r_i , whether legitimate or adversarial, first receives a nominal detector output $\mathcal{Z}_i^{\text{nat}}(t)$. Let $\lambda_{\text{FN},i}^s$ and $\lambda_{\text{FP},i}^s$ denote its nominal sensor false-negative and false-positive rates, respectively. For each visible object $o \in \mathcal{O}_i(t)$, the detector produces a noisy measurement $(id(o), \hat{\mathbf{p}}_{i,o}(t))$ with probability $1 - \lambda_{\text{FN},i}^s$, where $\hat{\mathbf{p}}_{i,o}(t) = \mathbf{p}_o(t) + \epsilon_{i,o}(t)$. The detector may also produce persistent sensor artifacts $\mathcal{Z}_i^s(t)$ with false-positive rate $\lambda_{\text{FP},i}^s$, giving

$$\mathcal{Z}_i^{\text{nat}}(t) = \mathcal{Z}_i^{\text{gt}}(t) \cup \mathcal{Z}_i^s(t). \quad (3)$$

Here $\mathcal{Z}_i^{\text{gt}}(t)$ collects the true-positive detections of visible objects retained with probability $1 - \lambda_{\text{FN},i}^s$, and $\mathcal{Z}_i^s(t)$ collects nominal sensor false positives.

Robots communicate with neighbors $\mathcal{N}_i(t) = \{r_j \neq r_i \mid \|\mathbf{p}_i(t) - \mathbf{p}_j(t)\| \leq R_{\text{com}}\}$ within the communication range R_{com} . Each robot forms its reported track set and broadcasts

$$\mu_i(t) = (\mathbf{s}_i(t), \mathcal{Z}_i(t)), \quad \mathcal{Z}_i(t) = \{\tau_i^k(t)\}_{k=1}^{N_i(t)}, \quad (4)$$

where

$$\tau_i^k(t) = (id_i^k, \hat{\mathbf{p}}_i^k(t)) \quad (5)$$

denotes the k -th local track reported by robot r_i , consisting of an object identifier and its estimated position. Here, id_i^k denotes a persistent track label supplied by the perception and association front end, rather than a ground-truth object identity available to the trust inference module. We assume that such labels can be matched across robot reports, while allowing residual association errors, including missed, incorrect, or duplicate associations. A track is active at time t if it is included in $\mathcal{Z}_i(t)$.

B. Reporting Policies

The robots are partitioned as $\mathcal{R} = \mathcal{R}^{\text{leg}} \cup \mathcal{R}^{\text{adv}}$, where the robot type is unknown to the trust inference algorithm. A legitimate robot does not manipulate its detector output:

$$\mathcal{Z}_i(t) = \mathcal{Z}_i^{\text{nat}}(t), \quad r_i \in \mathcal{R}^{\text{leg}}. \quad (6)$$

Following the perception-level attacks of [19], [20], adversarial robots receive nominal outputs from the same sensing model but manipulate which hypotheses they report, while their broadcast states $\mathbf{s}_i(t)$ remain authentic, e.g., verifiable through inter-robot localization [15].

Let $\mathcal{F}$ denote the set of persistent adversarial false hypotheses, with $|\mathcal{F}| \approx \lambda_{\text{FP}}^a |\mathcal{O}|$ and identifiers disjoint from $\mathcal{I}_{\text{gt}}$. Let $\mathcal{F}_i(t) \subseteq \mathcal{F}$ be the false hypotheses in r_i 's FoV. Consequently, multiple adversarial robots may report and mutually support the same persistent false hypothesis. These induce adversarial false-positive track observations $\mathcal{Z}_i^{\text{advFP}}(t) = \{(id(f), \hat{\mathbf{p}}_f(t)) \mid f \in \mathcal{F}_i(t)\}$ while suppressed nominal detections form $\mathcal{Z}_i^{\text{advFN}}(t) \subseteq \mathcal{Z}_i^{\text{nat}}(t)$.

In this work, we consider two standard adversarial policies. Under the *normal adversarial policy*, robot r_i randomly suppresses nominal detections with probability λ_{FN}^a and injects visible persistent false hypotheses:

$$\mathcal{Z}_i(t) = (\mathcal{Z}_i^{\text{nat}}(t) \setminus \mathcal{Z}_i^{\text{advFN}}(t)) \cup \mathcal{Z}_i^{\text{advFP}}(t), \quad r_i \in \mathcal{R}^{\text{adv}}. \quad (7)$$

Under the *optimized adversarial policy*, r_i selects reports in

$$\mathcal{C}_i(t) = \mathcal{Z}_i^{\text{nat}}(t) \cup \mathcal{Z}_i^{\text{advFP}}(t), \quad (8)$$

where $\mathcal{Z}_i^{\text{nat}}(t)$ contains nominal detector hypotheses and $\mathcal{Z}_i^{\text{advFP}}(t)$ contains persistent adversarial false hypotheses visible to r_i . For each $h \in \mathcal{C}_i(t)$, let $x_h \in \{0, 1\}$ indicate whether h is reported. Let $\mathcal{Y}_j(t) = \{id \mid (id, \hat{\mathbf{p}}) \in \mathcal{Z}_j(t)\}$ denote the identifiers reported by neighbor r_j . Using current neighbor states and one-step delayed reports, r_i computes

$$n_i^+(h, t) = \sum_{r_j \in \mathcal{N}_i(t)} \mathbf{1}[\hat{\mathbf{p}}_h(t) \in A_j(t)] \mathbf{1}[id(h) \in \mathcal{Y}_j(t-1)],$$

$$n_i^-(h, t) = \sum_{r_j \in \mathcal{N}_i(t)} \mathbf{1}[\hat{\mathbf{p}}_h(t) \in A_j(t)] \mathbf{1}[id(h) \notin \mathcal{Y}_j(t-1)].$$

The optimized adversary then solves

$$\begin{aligned} \max_{\{x_h\}_{h \in \mathcal{C}_i(t)}} \quad & \sum_{h \in \mathcal{Z}_i^{\text{advFP}}(t)} x_h (\delta_+ (1 + n_i^+(h, t)) - n_i^-(h, t)) + \\ & \sum_{h \in \mathcal{Z}_i^{\text{nat}}(t)} (1 - x_h) (\delta_- (1 + n_i^-(h, t)) - n_i^+(h, t)) \\ \text{s.t.} \quad & \sum_{h \in \mathcal{Z}_i^{\text{advFP}}(t)} x_h + \sum_{h \in \mathcal{Z}_i^{\text{nat}}(t)} (1 - x_h) \leq B_i, \\ & x_h \in \{0, 1\}, \quad \forall h \in \mathcal{C}_i(t). \end{aligned} \quad (9)$$

Here δ_+ weights the benefit of promoting adversarial false hypotheses, δ_- weights the benefit of suppressing nominal hypotheses, and B_i limits the number of manipulations per step for stealthiness. The reported set is

$$\mathcal{Z}_i(t) = \{h \in \mathcal{C}_i(t) \mid x_h = 1\} \quad r_i \in \mathcal{R}^{\text{adv}}. \quad (10)$$

C. Trust Inference Objective

A central trust module estimates a robot trust score $\gamma_i(t) \in [0, 1]$ for each r_i , while each robot r_i maintains track trust scores $\kappa_i^k(t) \in [0, 1]$ for its local tracks $\tau_i^k(t)$. These scores are inferred at deployment from cross-robot observation consistency, without requiring prior trust, reputation, or ground-truth labels at inference time.

Given threshold ψ , the inferred legitimate robot set is

$$\hat{\mathcal{R}}^{\text{leg}}(T) = \{r_i \in \mathcal{R} \mid \gamma_i(T) \geq \psi\}. \quad (11)$$

For evaluation, local tracks are grouped into fused hypotheses $\hat{\mathcal{O}}_{\text{all}}$. For each fused hypothesis $\bar{o}$, let $\mathcal{P}_{\bar{o}}$ be the set of reporting robot-track pairs. Its credibility is

$$\Gamma_{\bar{o}}(T) = \frac{\sum_{(i,k) \in \mathcal{P}_{\bar{o}}: \gamma_i(T) \geq \psi} \gamma_i(T) \kappa_i^k(T)}{\sum_{(i,k) \in \mathcal{P}_{\bar{o}}: \gamma_i(T) \geq \psi} \gamma_i(T)}, \quad (12)$$

and zero if the denominator is zero. The credible object set is

$$\hat{\mathcal{O}}_{\text{fused}}(T) = \{\bar{o} \in \hat{\mathcal{O}}_{\text{all}} \mid \Gamma_{\bar{o}}(T) \geq \psi\}. \quad (13)$$

The objective is to identify legitimate robots and credible object hypotheses despite natural sensor artifacts and adversarial false hypotheses.

IV. NEST-BAYES: NEURAL-SYMBOLIC TRUST WITH BAYESIAN EVIDENCE ACCUMULATION

We introduce *NeST-Bayes*, a trust inference framework based on ego-centric relational graphs encoding cross-robot observation consistency (Fig. 1). At each step, a centralized Trust Aggregation Module (TAM) constructs $\mathcal{G}_i(t)$ for each robot r_i from r_i 's own and neighboring reports and applies a shared neural-symbolic model to reason through each graph, accumulating ego-robot evidence in a system-wide registry while returning validated track evidence to r_i , whose track beliefs are maintained onboard. The TAM could equally be replicated onboard each robot; we centralize it to obtain a single system-level trust estimate per robot, leaving the full decentralized deployment to future work.

A. Single Step Neural-Symbolic Reasoning

1) *Ego-Centric Graph Construction*: At each time step t , the TAM constructs for robot r_i a directed, heterogeneous ego-centric graph $\mathcal{G}_i(t) = (\mathcal{V}_i(t), \mathcal{E}_i(t))$ from r_i 's message $\mu_i(t)$ and the messages $\{\mu_j(t)\}_{r_j \in \mathcal{N}_i(t)}$ of its neighbors, where each message contains the robot state and reported tracks. The graph translates the reported observations into a symbolic topology that facilitates trust inference.

$$\mathcal{V}_i(t) = \{r_i\} \cup \mathcal{N}_i(t) \cup \tilde{\mathcal{T}}_i^{\text{fused}}(t), \quad (14)$$

where $\tilde{\mathcal{T}}_i^{\text{fused}}(t)$ is obtained by grouping the tracks reported by $\{r_i\} \cup \mathcal{N}_i(t)$ according to their persistent identifiers. Let $\mathcal{I}_i(t)$ denote the set of identifiers appearing in these reports. For each $id \in \mathcal{I}_i(t)$, define

$$\mathcal{R}_i^{\text{id}}(t) = \{r_j \in \{r_i\} \cup \mathcal{N}_i(t) \mid \exists \tau_j^k(t) = (id, \hat{\mathbf{p}}_j^k(t)) \in \mathcal{Z}_j(t)\},$$

$$\tilde{\tau}_i^{\text{id}}(t) = \left(id, \frac{1}{|\mathcal{R}_i^{\text{id}}(t)|} \sum_{r_j \in \mathcal{R}_i^{\text{id}}(t)} \hat{\mathbf{p}}_j^{\text{id}}(t), \mathcal{R}_i^{\text{id}}(t) \right),$$

where $\hat{\mathbf{p}}_j^{\text{id}}(t)$ is the position estimate associated with id in robot r_j 's report. The fused track set is $\tilde{\mathcal{T}}_i^{\text{fused}}(t) = \{\tilde{\tau}_i^{\text{id}}(t) \mid id \in \mathcal{I}_i(t)\}$, where id is the persistent track label of Sec. III; residual missed, incorrect, or duplicate associations thus manifest as track-level false negatives or positives, so *NeST-Bayes* operates on potentially imperfect associations.

TABLE I
SYMBOLIC EDGE SEMANTICS IN THE EGO-CENTRIC GRAPH

Relation Category	Edge Type	Condition & Interpretation
Robot-Track Relations (Sensing Consistency)		
Observation	InFoV_Observed	A robot reports observing a fused track within its FoV.
Refutation	InFoV_Missed	A robot fails to observe a fused track within its FoV.
Robot-Robot Relations (Cross-Robot Consistency)		
Consistency	Co-Detected	Both robots report observation of a fused track within their FoV.
Conflict	Contradicted	A robot reports observation that another robot refutes within its FoV.

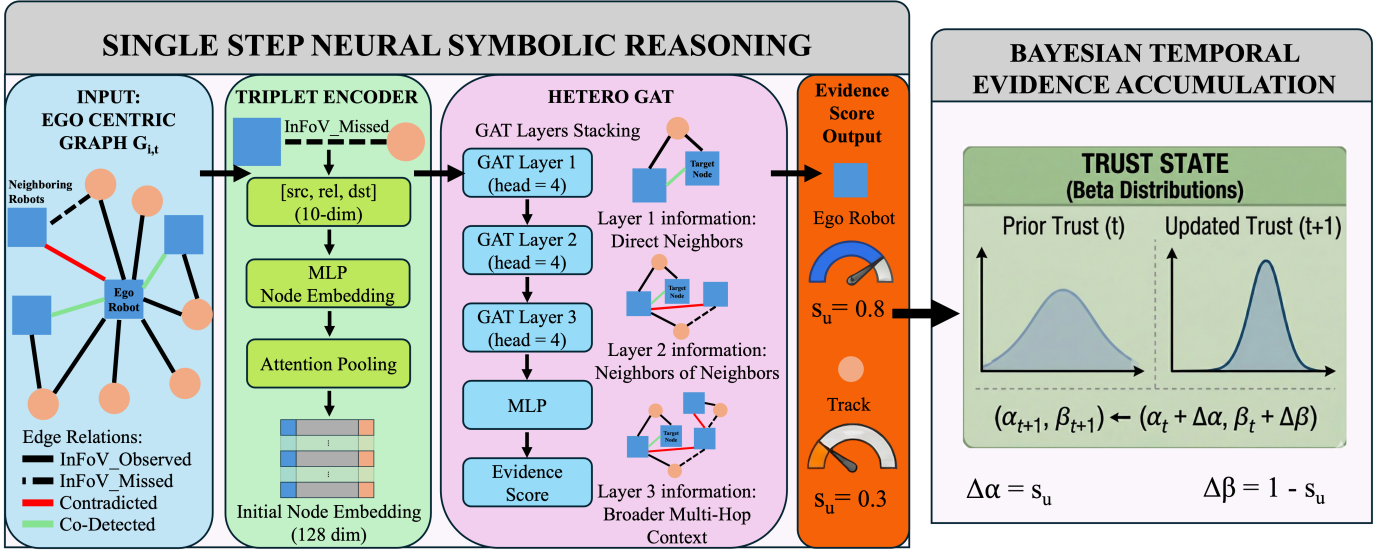


Fig. 1. *NeST-Bayes* pipeline: ego-centric graph construction, neural-symbolic evidence inference, and Bayesian evidence accumulation.

The edges $\mathcal{E}_i(t)$ represent specific symbolic relations derived from geometric consistency checks and FoV analysis. These relations encode the logical “evidence” for or against a node’s trustworthiness. To enable bidirectional information flow, in which robots validate tracks and tracks reciprocally validate robots, we model each relationship as a pair of directed edges (forward and reverse). This yields eight directed edge types, with the four forward relations defined in Table I. The TAM constructs this graph (Alg. 1, lines 1–7) and then performs inference on it.

2) *Neural-Symbolic Reasoning*: This stage maps the symbolic abstraction of Sec. IV-A into instantaneous evidence scores for both robots and object hypotheses, which are subsequently accumulated via Beta updates (Sec. IV-B). Rather than applying fixed aggregation rules to individual relations, we cast trust inference as typed relational reasoning over the ego-centric graph. Geometric and FoV consistency checks provide interpretable symbolic predicates, while stacked relation-specific message passing learns how local predicates combine into broader multi-hop evidence. This construction is inspired by neural-logic relational reasoning, but does not instantiate the explicit predicate operators of an NLM.

In node feature initialization, as in the Triplet Encoder section in Fig. 1, the semantic profile of a node is defined by the set of symbolic relations it participates in. For each node $u \in \mathcal{V}_i(t)$, we represent each relation from node u to any other node v as a symbolic triplet:

$$\eta(u, v) = [\phi(u), e(\rho), \phi(v)] \in \{0, 1\}^{10} \quad (15)$$

where $\phi(\cdot)$ is a one-bit node-type indicator (robot or track), and $e(\rho)$ is an 8-bit one-hot encoding of the relation type ρ defined in Table I and their reverses. These triplets form an unordered bag of symbolic predicates, $\mathcal{H}_u$, that characterize the local consistency context of a node u .

Let $c = \text{type}(u) \in \{\text{robot}, \text{track}\}$. Each triplet $\eta \in \mathcal{H}_u$ is embedded by a type-specific two-layer MLP g_c and pooled

Algorithm 1 Single Step Neural-Symbolic Reasoning

Require: Neighbor messages $\mu_j(t)$ for all $r_j \in \{r_i\} \cup \mathcal{N}_i(t)$

Ensure: Validated evidence scores $\Omega_i = \{(u, s_u)\}$

Phase 1: Ego-Centric Graph Construction

- 1: $\mathcal{V}_i(t) \leftarrow \{r_i\} \cup \mathcal{N}_i(t) \cup \tilde{\mathcal{T}}_i^{\text{fused}}(t)$, $\mathcal{E}_i(t) \leftarrow \emptyset$
- 2: **for** each $(u, v) \in \mathcal{V}_i(t) \times \mathcal{V}_i(t)$ **do**
- 3: **if** relation ρ holds (Table I) **then**
- 4: $\mathcal{E}_i(t) \leftarrow \mathcal{E}_i(t) \cup \{(u, \rho, v), (v, \rho^{-1}, u)\}$
- 5: **end if**
- 6: **end for**

Phase 2: Neural-Symbolic Reasoning

- 7: $\mathcal{H} \leftarrow \text{ExtractTriplets}(\mathcal{E}_i(t))$
- 8: $\mathbf{h}^{(0)} \leftarrow \text{AttentionPooling}(\mathcal{H})$
- 9: $\mathbf{h}^{(1)} \leftarrow \text{HeteroGAT}^{(1)}(\mathbf{h}^{(0)}, \mathcal{E}_i(t))$
- 10: **for** $\ell = 2$ to L **do**
- 11: $\hat{\mathbf{h}}^{(\ell)} \leftarrow \text{HeteroGAT}^{(\ell)}(\mathbf{h}^{(\ell-1)}, \mathcal{E}_i(t))$
- 12: $\mathbf{h}^{(\ell)} \leftarrow \hat{\mathbf{h}}^{(\ell)} + \mathbf{h}^{(\ell-1)}$
- 13: **end for**

Phase 3: Validity Filtering

- 14: $\Omega_i \leftarrow \emptyset$
- 15: **for** each eligible node u **do**
- 16: **if** $\text{IsValidNode}(u, \mathcal{E}_i(t))$ **then**
- 17: $s_u \leftarrow \sigma(\text{MLP}_{\text{type}(u)}(\mathbf{h}_u^{(L)}))$
- 18: $\Omega_i \leftarrow \Omega_i \cup \{(u, s_u)\}$
- 19: **end if**
- 20: **end for**
- 21: **return** Ω_i ▷ Output evidence scores

according to

$$a_\eta = \mathbf{w}_c^\top \tanh(W_c g_c(\eta)), \quad \pi_\eta = \frac{\exp(a_\eta)}{\sum_{\eta' \in \mathcal{H}_u} \exp(a_{\eta'})},$$

$$\mathbf{h}_u^{(0)} = W_{o,c} \sum_{\eta \in \mathcal{H}_u} \pi_\eta g_c(\eta). \quad (16)$$

This permutation invariant pooling produces a 128 dimen-

sional initial embedding for the robot and track nodes.

The embeddings are processed by $L = 3$ relation-specific Hetero GAT layers. Each layer applies standard graph attention independently over each directed relation type using $K = 4$ heads. Head outputs and incoming relation representations are mean-aggregated, followed by ReLU and type-specific batch normalization. Residual connections are applied after the second and third layers, allowing evidence to propagate from local consistency relations to broader multi-hop consensus.

The final node embeddings, $\mathbf{h}_u^{(L)}$, are passed to type-specific MLP heads with a Sigmoid activation function $\sigma(\cdot)$ to produce scalar evidence scores $s_u \in [0, 1]$.

$$s_u = \sigma(\text{MLP}_{\text{type}(u)}(\mathbf{h}_u^{(L)})) \quad (17)$$

3) *Validity Filtering*: The output of neural-symbolic reasoning is filtered before the trust update. Validity is defined differently for robots and tracks:

- **Robots**: The ego robot r_i is considered valid for a robot trust update only if it has at least one incident edge involving a neighboring robot in $\mathcal{N}_i(t)$.
- **Tracks**: A local track $\tau_i^k(t) \in \mathcal{Z}_i(t)$ is considered valid for a track trust update only if its corresponding fused track node is connected to the ego robot and at least one neighboring robot in $\mathcal{N}_i(t)$.

Although the GNN scores all nodes in $\mathcal{G}_i(t)$, only the valid ego-robot score and r_i 's local-track scores are retained. The TAM applies the ego score to r_i 's system-level trust belief, ensuring at most one robot update per step, and maps each valid fused-track score by its persistent identifier back to the local track $\tau_i^k(t)$ for r_i 's onboard update.

B. Bayesian Temporal Evidence Accumulation

To accumulate instantaneous neural evidence over time, we maintain a Bayesian trust belief for each node u , represented by a Beta distribution over its trustworthiness $\xi_u \in [0, 1]$:

$$f(\xi_u; \alpha_u(t), \beta_u(t)) = \frac{\xi_u^{\alpha_u(t)-1} (1-\xi_u)^{\beta_u(t)-1}}{\text{Beta}(\alpha_u(t), \beta_u(t))} \quad (18)$$

where $\alpha_u(t)$ represents the accumulated positive evidence and $\beta_u(t)$ represents the accumulated negative evidence up to time t . Accordingly, the TAM maintains (α_i, β_i) in a system-wide robot-trust registry, while each robot r_i maintains (α_i^k, β_i^k) for its own local tracks. This division provides a globally consistent robot-trust estimate while keeping track beliefs colocated with the local perception and planning modules that consume them. Each newly encountered robot or track is initialized with the symmetric prior $(\alpha_0, \beta_0) = (1, 1)$, corresponding to a neutral trust score of 0.5. Entities receiving no valid cross-robot evidence retain this prior.

The Hetero GAT output score $s_u \in [0, 1]$ is interpreted as the instantaneous evidence score. To ensure that evidence reinforces the correct belief parameter, we accumulate it as fractional pseudo-counts in the style of Beta-reputation and subjective-logic systems [45]:

$$\alpha_u(t) = \alpha_u(t-1) + s_u, \quad \beta_u(t) = \beta_u(t-1) + (1 - s_u) \quad (19)$$

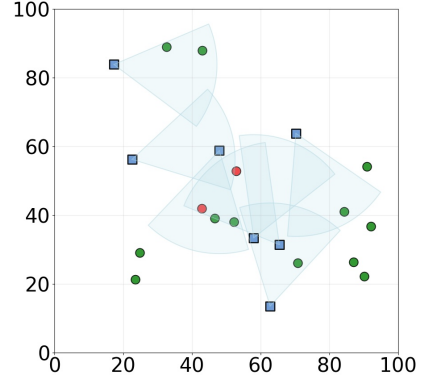


Fig. 2. Example simulation environment. Robots are shown as blue squares with shaded fields of view, ground-truth objects as green points, and false positive objects as red points.

This update is only applied to nodes that satisfy the validity constraints defined in Alg. 1. For non-valid nodes, the α and β parameters remain unchanged.

The final trust score $\hat{\gamma}_u(t)$ used for decision-making is the expected value of the Beta belief:

$$\hat{\gamma}_u(t) = \mathbb{E}[\xi_u] = \frac{\alpha_u(t)}{\alpha_u(t) + \beta_u(t)} \quad (20)$$

Over the horizon, the TAM repeats single-step reasoning for every robot, applies the update in Eq. (19) to each validated evidence pair, and reports the final robot and track trust scores as the Beta expectations in Eq. (20), i.e., $\gamma_i(T)$ and $\kappa_i^k(T)$ in Eqs. (11)–(12). This ensures trust reflects accumulated evidence rather than any single snapshot.

V. EVALUATIONS

We first train and validate *NeST-Bayes* in a controlled 100×100 2D environment (Fig. 2). Each robot has an occlusion-free FoV of $\pi/3$ and 50-unit range, with 80-unit communication range; we fix $|\mathcal{O}| = 2|\mathcal{R}|$ and sample $|\mathcal{R}| \in [5, 20]$. Robots and objects follow random start-goal motions over 100 steps. During both training and testing, nominal sensor false-positive and false-negative rates are independently sampled from $[0.01, 0.05]$ for every robot, including legitimate and adversarial robots. Training samples the adversarial ratio $m \in [0.1, 0.3]$ and, for normal adversaries, $\lambda_{\text{FP}}^a, \lambda_{\text{FN}}^a \in [0.1, 0.3]$, with scenarios split evenly between normal and optimized aggressive policies. From 10,000 scenarios we extract 71,940 valid ego-centric graphs, partitioned 80/20 by scenario into train/validation. The model is supervised only on valid nodes (Sec. IV-A3) via:

$$\mathcal{L}_i = 5 \sum_{r \in \mathcal{V}_{R,i}^{\text{tr}}} \ell_{\text{BCE}}(y_r, s_r) + \sum_{\tilde{r} \in \mathcal{V}_{T,i}^{\text{tr}}} \ell_{\text{BCE}}(y_{\tilde{r}}, s_{\tilde{r}}), \quad (21)$$

where $y = 1$ marks legitimate robots and real hypotheses; the factor 5 balances the single ego-robot label against multiple track labels, and the Beta updates add no learned parameters.

A. Baselines and Evaluation Metrics

In addition to a *No Trust* baseline, which indiscriminately trusts all robots and therefore propagates adversarial observations, we also compare the proposed *NeST-Bayes* method with

TABLE II

PRECISION AND RECALL UNDER NORMAL AND OPTIMIZED ADVERSARIAL POLICIES. EACH ENTRY REPORTS PRECISION $\pm$ STANDARD DEVIATION FOLLOWED BY RECALL $\pm$ STANDARD DEVIATION, IN PERCENT. EACH SETTING IS EVALUATED OVER 100 INDEPENDENTLY GENERATED SCENARIOS.

Setting	Entity	Precision $\pm$ Std./Recall $\pm$ Std. (%)			
		No Trust	Naïve Bayesian	MATE	NeST-Bayes
Normal Adversarial Policy					
Nominal	Robot	85.3 $\pm$ 5.5/100.0 $\pm$ 0.0	93.3 $\pm$ 8.8/79.2 $\pm$ 20.6	95.0 $\pm$ 7.8/84.4 $\pm$ 17.2	98.6$\pm$3.7/95.8$\pm$8.2
	Object	83.3 $\pm$ 5.5/100.0 $\pm$ 0.0	94.7 $\pm$ 7.4/95.1 $\pm$ 10.9	95.0 $\pm$ 7.5/87.1 $\pm$ 11.4	98.0$\pm$5.0/93.7$\pm$8.4
High FP Injection ($\lambda_{FP}^a \uparrow$)	Robot	85.3 $\pm$ 5.5/100.0 $\pm$ 0.0	95.7 $\pm$ 8.0/77.1 $\pm$ 22.1	95.3 $\pm$ 7.5/84.3 $\pm$ 17.2	99.0$\pm$3.4/94.9$\pm$9.7
	Object	67.1 $\pm$ 4.3/100.0 $\pm$ 0.0	89.1 $\pm$ 13.6/94.5 $\pm$ 11.7	90.9 $\pm$ 13.7/86.7 $\pm$ 11.1	96.9$\pm$9.6/93.6$\pm$8.4
High FN Suppression ($\lambda_{FN}^a \uparrow$)	Robot	85.3 $\pm$ 5.5/100.0 $\pm$ 0.0	97.4 $\pm$ 14.2/75.5 $\pm$ 25.1	97.9 $\pm$ 5.3/81.5 $\pm$ 20.2	99.4$\pm$2.2/94.3$\pm$10.7
	Object	83.4 $\pm$ 5.2/100.0 $\pm$ 0.0	95.5 $\pm$ 7.3/91.5 $\pm$ 16.2	95.7 $\pm$ 7.2/80.5 $\pm$ 14.1	99.1$\pm$3.3/93.3$\pm$8.7
Optimized Adversarial Policy					
Conservative Policy	Robot	85.1 $\pm$ 5.6/100.0 $\pm$ 0.0	88.8 $\pm$ 8.1/86.9 $\pm$ 17.0	83.8 $\pm$ 9.1/83.3 $\pm$ 18.9	86.5$\pm$7.6/92.5$\pm$11.4
	Object	88.3 $\pm$ 7.4/100.0 $\pm$ 0.0	95.5 $\pm$ 5.9/94.5 $\pm$ 12.1	96.3 $\pm$ 5.4/87.9 $\pm$ 11.3	97.0$\pm$5.1/94.4$\pm$7.4
Moderate Policy ($\delta_+, \delta_- \uparrow$)	Robot	85.1 $\pm$ 5.6/100.0 $\pm$ 0.0	93.2 $\pm$ 6.7/83.1 $\pm$ 19.0	90.8 $\pm$ 9.0/80.2 $\pm$ 18.3	96.0$\pm$5.9/92.9$\pm$10.4
	Object	87.1 $\pm$ 7.2/100.0 $\pm$ 0.0	95.8 $\pm$ 5.5/93.5 $\pm$ 13.1	97.8 $\pm$ 4.6/83.5 $\pm$ 13.5	97.9$\pm$4.2/94.0$\pm$7.9
Aggressive Policy ($\delta_+, \delta_- \uparrow\uparrow$)	Robot	85.1 $\pm$ 5.6/100.0 $\pm$ 0.0	96.5 $\pm$ 5.3/79.9 $\pm$ 21.0	93.6 $\pm$ 8.6/80.7 $\pm$ 18.2	98.8$\pm$4.2/93.8$\pm$10.4
	Object	86.1 $\pm$ 6.9/100.0 $\pm$ 0.0	97.1 $\pm$ 5.2/93.5 $\pm$ 13.1	97.6 $\pm$ 4.5/83.7 $\pm$ 13.5	98.3$\pm$3.7/95.4$\pm$6.7
High Adv. Ratio ($m \in [0.4, 0.5]$)	Robot	62.1 $\pm$ 4.0/100.0 $\pm$ 0.0	73.8 $\pm$ 34.9/49.3 $\pm$ 30.4	58.9 $\pm$ 25.4/59.0 $\pm$ 27.8	89.1$\pm$15.2/85.9$\pm$16.2
	Object	80.6 $\pm$ 5.6/100.0 $\pm$ 0.0	86.2 $\pm$ 21.2/82.2 $\pm$ 23.8	81.5 $\pm$ 21.4/67.4 $\pm$ 23.6	93.3$\pm$8.3/86.5$\pm$13.3

two representative trust aggregation approaches commonly used in multi-robot systems: (i) *Naïve Bayesian*, similar to [15], that maintains independent trust estimates for robot and track nodes by directly accumulating positive and negative evidence over time to update the beta trust distribution: Co-Detected and InFoV_Observed edges are treated as positive evidence for robot and track nodes respectively, while Contradicted and InFoV_Missed edges are treated as negative evidence. Importantly, this baseline consumes exactly the same symbolic relations as *NeST-Bayes* but weights them by fixed rules and treats each robot and track as an independent evidence accumulator. Any gap therefore reflects how the evidence is interpreted rather than the evidence itself; (ii) *MATE* [20] is a rule-based trust estimation method that maps perception consistency checks into trust pseudomeasurements (PSMs) and updates trust scores for its underlying beta distributions through hand-designed aggregation rules. In all tests, trust is accumulated over a horizon of 100 time steps. Final classifications use the fixed threshold $\psi = 0.5$. We report precision and recall for both robot and object classification to separately measure the reliability of the retained information and the coverage of legitimate robots and ground-truth objects. Precision is averaged over scenarios and scored 0 for both robots and objects when a method retains no robot.

B. Generalization with Normal Adversaries

We evaluate normal adversaries that consistently report visible false-positive hypotheses while suppressing nominal detections with probability λ_{FP}^a . Each setting uses 100 scenarios; results are in Table II. Across all three settings, *NeST-Bayes* achieves the highest mean robot precision (98.6–99.4%) and recall (94.3–95.8%) among trust-based methods, together with the highest object precision (96.9–99.1%). Under high false-positive injection ($\lambda_{FP}^a \in [0.4, 0.8]$), it achieves 96.9%/93.6% object precision/recall, compared with 89.1%/94.5% for *Naïve Bayesian* and 90.9%/86.7% for *MATE*. Under high false-negative suppression ($\lambda_{FN}^a \in [0.4, 0.8]$), *NeST-Bayes* leads in

both precision and recall for both entity types. The comparison with *Naïve Bayesian*, which accumulates identical symbolic evidence, demonstrates the value of learned relational reasoning for both the reliability and coverage of retained information (Sec. V-D).

C. Generalization with Optimized Adversaries

We next evaluate adversaries that optimize both the timing and type of their manipulations via Eq. (9), with $\delta_+ = \delta_- \in \{1, 2, 3\}$ yielding conservative, moderate, and aggressive policies and the attack budget fixed at $B_i = 5$. More aggressive policies are preferable when the adversary seeks to maximally disrupt neighbors’ motion planning: false positives divert robots toward nonexistent obstacles, while suppressed detections conceal real hazards. Results are shown in Table II.

The conservative policy selectively injects false hypotheses that neighbors already appear to support and suppresses nominal detections lacking neighbor corroboration, producing few observable inconsistencies. Across trust-based methods, robot precision remains close to the 85.1% *No Trust* baseline (83.8–88.8%), indicating limited robot-level discrimination in this setting. *NeST-Bayes* achieves 86.5% robot precision with 92.5% recall. At the object level, it increases precision from 88.3% for *No Trust* to 97.0%, while retaining 94.4% recall. This is the highest mean object precision among trust-based methods, compared with 95.5% for *Naïve Bayesian* and 96.3% for *MATE*. Its object recall is comparable to *Naïve Bayesian*’s 94.5% and exceeds *MATE*’s 87.9%. These results show that object-hypothesis validation remains effective when robot-level discrimination is limited.

As δ_+ and δ_- increase, attacks become more detectable through cross-robot disagreement. *NeST-Bayes* robot precision rises from 86.5% under conservative attacks to 98.8% under aggressive attacks, a 12.3-percentage-point gain compared with 7.7 points for *Naïve Bayesian* (88.8% to 96.5%). Object precision increases from 97.0% to 98.3%. Under aggressive attacks, *NeST-Bayes* also achieves the highest robot and object

TABLE III
COMPONENT ABLATION OVER 100 COMPREHENSIVE SCENARIOS.
RESULTS ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION.

Method	Precision / Recall (%)	
	Robot	Object
<i>No Trust</i>	84.7 $\pm$ 5.4 / 100.0 $\pm$ 0.0	86.2 $\pm$ 5.4 / 100.0 $\pm$ 0.0
<i>No Hetero GAT</i>	95.0 $\pm$ 11.7 / 75.4 $\pm$ 26.1	98.5 $\pm$ 3.6 / 89.3 $\pm$ 12.6
<i>Homogeneous GAT</i>	86.8 $\pm$ 12.7 / 78.7 $\pm$ 23.2	87.0 $\pm$ 5.8 / 96.1 $\pm$ 6.6
<i>Type Embeddings Only</i>	95.8 $\pm$ 7.1 / 91.3 $\pm$ 10.7	98.2 $\pm$ 4.0 / 94.4 $\pm$ 7.1
NeST-Bayes (Full)	97.2$\pm$6.2 / 95.5$\pm$8.8	99.0$\pm$2.9 / 94.8$\pm$7.8

recall among trust-based methods (93.8% and 95.4%). With the adversarial ratio increased to $m \in [0.4, 0.5]$, *NeST-Bayes* leads in all four metrics: 89.1% robot precision with 85.9% recall and 93.3% object precision with 86.5% recall. Compared with *Naïve Bayesian*, robot precision and recall improve by 15.3 and 36.6 percentage points, respectively. These joint gains demonstrate effective filtering and retention of legitimate information when adversaries are numerous.

D. Ablation Studies

We evaluate each ablation over 100 scenarios, evenly split between normal and optimized adversaries, with the latter equally sampling conservative, moderate, and aggressive policies. Table III shows that the full model achieves the highest mean robot precision and recall among learned variants (97.2% and 95.5%), together with the highest object precision (99.0%) and 94.8% object recall. Removing the Hetero GAT reduces the model to node-wise MLPs and lowers robot recall from 95.5% to 75.4% and object recall from 94.8% to 89.3%. These 20.1 and 5.5 percentage gains demonstrate the contribution of cross-node evidence propagation. Compared with a Homogeneous GAT, which collapses all relation types into a single edge type, typed relations improve robot precision and recall by 10.4 and 16.8 percentage points, respectively, and object precision by 12.0 points, showing the value of distinguishing agreement from contradiction. Replacing triplet-based initialization with learned node-type embeddings lowers robot precision from 97.2% to 95.8% and recall from 95.5% to 91.3%, demonstrating the benefit of encoding local symbolic relation profiles. Together, these results support the complementary contributions of symbolic initialization and heterogeneous relational reasoning.

E. High Fidelity Simulations

We also evaluate *NeST-Bayes* without retraining in a high-fidelity Webots [46] warehouse simulation with 5 Spot robots and 26 ground-truth objects, including dynamic (walking workers) and static objects (e.g., cardboard boxes falling from shelves), as in Fig. 3. Each robot uses two head-mounted cameras ($\pi/4$ FoV, 20 m range), with FoV and occlusion reasoning from a known warehouse layout used only for visibility checks. Observations follow the same nominal sensing-noise model as in the previous sections. In each of 100 scenarios, two robots follow an optimized policy sampled from the conservative, moderate, and aggressive settings of Sec. V-C. The results are reported in Table IV.



Fig. 3. Webots warehouse simulation environment with five Spot robots and a mixture of objects. Occlusions from shelves and dynamic human agents create partial observability and asymmetric sensing among robots.

TABLE IV
PRECISION AND RECALL IN THE WEBOTS SIMULATIONS OVER 100 SCENARIOS. RESULTS ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION.

Method	Precision / Recall (%)	
	Robot	Object
<i>No Trust</i>	60.0 $\pm$ 0.0 / 100.0 $\pm$ 0.0	71.4 $\pm$ 6.3 / 100.0 $\pm$ 0.0
<i>Naïve Bayesian</i>	87.7 $\pm$ 32.7 / 61.3 $\pm$ 33.7	87.7 $\pm$ 32.7 / 81.6 $\pm$ 31.5
<i>MATE</i>	67.0 $\pm$ 47.3 / 47.0 $\pm$ 40.8	67.0 $\pm$ 47.3 / 38.1 $\pm$ 28.8
NeST-Bayes (Ours)	92.5$\pm$13.1 / 92.0$\pm$14.3	97.3$\pm$5.4 / 85.0$\pm$7.5

Frequent occlusions, limited FoV overlap, and dynamic objects create sparse cross-robot validation in Webots. *NeST-Bayes* achieves the highest mean precision and recall for both entity types among trust-based methods: 92.5% robot precision with 92.0% recall, and 97.3% object precision with 85.0% recall. Relative to *Naïve Bayesian*, robot precision and recall improve by 4.8 and 30.7 percentage points, respectively, while object precision and recall improve by 9.6 and 3.4 points. *NeST-Bayes* also exhibits lower variability in robot precision (13.1 percentage points standard deviation) than *Naïve Bayesian* (32.7) and *MATE* (47.3). These results demonstrate effective transfer without retraining, improving both information reliability and coverage in an occlusion-rich environment.

VI. DISCUSSION AND CONCLUSION

This paper presented *NeST-Bayes*, a neural-symbolic framework that jointly infers robot and object-hypothesis trust from cross-robot observation consistency. By encoding observation agreement and contradiction as symbolic relations in ego-centric heterogeneous graphs and accumulating the resulting neural evidence through Bayesian updates of Beta-distributed trust beliefs, *NeST-Bayes* operates without prior trust assignments, side-channel information, or hand-designed nominal behavior specifications. Comparisons against *Naïve Bayesian*, which accumulates identical symbolic evidence, demonstrate the benefits of learned relational reasoning for both precision and recall, including under occlusion-rich conditions. Component ablations show that heterogeneous relational propagation improves the coverage of legitimate robots and real objects. Across the benchmark settings, *NeST-Bayes* achieves the highest mean object precision and robot recall among trust-based methods, while also leading in robot precision and object recall in most settings.

Future work will explore trust-aware active verification, directing robots to obtain additional viewpoints or revisit uncertain detections under sparse cross-robot evidence.

REFERENCES

- [1] W. Burgard, M. Moors, C. Stachniss, and F. E. Schneider, "Coordinated multi-robot exploration," *IEEE Transactions on robotics*, vol. 21, no. 3, pp. 376–386, 2005.
- [2] J. Cortés and M. Egerstedt, "Coordinated control of multi-robot systems: A survey," *SICE Journal of Control, Measurement, and System Integration*, vol. 10, no. 6, pp. 495–503, 2017.
- [3] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-robot collaborative perception with graph neural networks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2289–2296, 2022.
- [4] R. Wang, H.-L. Hsu, D. Hunt, J. Kim, S. Luo, and M. Pajic, "Comres-rlm: Coordinated multi-robot exploration and search using vision language models," *arXiv preprint arXiv:2509.26324*, 2025.
- [5] Z. Xu, X. Lin, and V. Tzoumas, "Bandit submodular maximization for multi-robot coordination in unpredictable and partially observable environments," in *Robotics: Science and Systems*, 2023.
- [6] A. Mannucci, L. Pallottino, and F. Pecora, "Provably safe multi-robot coordination with unreliable communication," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3232–3239, 2019.
- [7] M. Catellani and L. Sabattini, "Distributed control of a limited angular field-of-view multi-robot system in communication-denied scenarios: A probabilistic approach," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 739–746, 2023.
- [8] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, "Guaranteeing spoof-resilient multi-robot networks," *Autonomous Robots*, vol. 41, no. 6, pp. 1383–1400, 2017.
- [9] J. Li, W. Abbas, M. Shabbir, and X. D. Koutsoukos, "Resilient distributed diffusion for multi-robot systems using centerpoint," in *Robotics: science and systems*, vol. 2020. Corvallis, OR, 2020.
- [10] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Resilient distributed state estimation with mobile agents: overcoming byzantine adversaries, communication losses, and intermittent measurements," *Autonomous Robots*, vol. 43, no. 3, pp. 743–768, 2019.
- [11] C. Burnett, T. J. Norman, and K. Sycara, "Trust decision-making in multi-agent systems," in *Proceedings of the twenty-second international joint conference on artificial intelligence-volume one*, 2011, pp. 115–120.
- [12] Y. Yao, H. Tong, X. Yan, F. Xu, and J. Lu, "Matri: a multi-aspect and transitive trust inference model," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1467–1476.
- [13] S. Asiri and A. Miri, "An iot trust and reputation model based on recommender systems," in *2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 2016, pp. 561–568.
- [14] A. Pierson and M. Schwager, "Adaptive inter-robot trust for robust multi-robot sensor coverage," in *Robotics research: The 16th international symposium ISRR*. Springer, 2016, pp. 167–183.
- [15] M. Cheng, C. Yin, J. Zhang, S. Nazarian, J. Deshmukh, and P. Bogdan, "A general trust framework for multi-agent systems," in *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems*, 2021, pp. 332–340.
- [16] M. Cavorsi, F. Mallmann-Trenn, D. Saldana, and S. Gil, "Dynamic crowd vetting: Collaborative detection of malicious robots in dynamic communication networks," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 7546–7553.
- [17] M. Cavorsi, O. E. Akgün, M. Yemini, A. J. Goldsmith, and S. Gil, "Exploiting trust for resilient hypothesis testing with malicious robots," *IEEE Transactions on Robotics*, vol. 40, pp. 3514–3536, 2024.
- [18] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, 2022.
- [19] R. Bahrami and H. Jafarnejadsani, "Multi-robot coordination with adversarial perception," in *2025 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2025, pp. 370–377.
- [20] R. S. Hallyburton and M. Pajic, "Security-aware sensor fusion with mate: the multi-agent trust estimator," in *Proc. the 2025 ACM SIGSAC Conf. on Computer and Communications Security*, 2025, pp. 2009–2023.
- [21] M. Lauri, D. Hsu, and J. Pajarinen, "Partially observable markov decision processes in robotics: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 21–40, 2022.
- [22] Y. Zhang and L. E. Parker, "Iq-asymtre: Forming executable coalitions for tightly coupled multirobot tasks," *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 400–416, 2012.
- [23] Y. Yan and Y. Mostofi, "Robotic router formation in realistic communication environments," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 810–827, 2012.
- [24] N. Bezzo, J. Weimer, M. Pajic, O. Sokolsky, G. J. Pappas, and I. Lee, "Attack resilient state estimation for autonomous robotic systems," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3692–3698.
- [25] D. Yeke, K. A. Pant, M. O. Ozmen, H. Kim, J. M. Goppert, I. Hwang, A. Bianchi, and Z. B. Celik, "Automated discovery of semantic attacks in {Multi-Robot} navigation systems," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 3959–3978.
- [26] G. Deng, Y. Zhou, Y. Xu, T. Zhang, and Y. Liu, "An investigation of byzantine threats in multi-robot systems," in *Proc. 24th int. symposium on research in attacks, intrusions and defenses*, 2021, pp. 17–32.
- [27] J. Li, W. Abbas, M. Shabbir, and X. Koutsoukos, "Byzantine resilient distributed learning in multirobot systems," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3550–3563, 2022.
- [28] Y. Guo, X. J. Yang, and C. Shi, "Enabling team of teams: A trust inference and propagation (tip) model in multi-human multi-robot teams," in *Robotics: science and systems*, 07 2023.
- [29] R. S. Hallyburton and M. Pajic, "Trust-based assured sensor fusion in distributed aerial autonomy," in *Proc. the ACM/IEEE 16th International Conference on Cyber-Physical Systems (ICCPs)*, 2025, pp. 1–12.
- [30] —, "Bayesian methods for trust in collaborative multi-agent autonomy," in *63rd IEEE Conference on Decision and Control (CDC)*, 2024, pp. 470–476.
- [31] T. R. Besold, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L. C. Lamb, P. M. V. Lima, L. de Penning, G. Pinkas *et al.*, "Neural-symbolic learning and reasoning: A survey and interpretation 1," in *Neuro-symbolic artificial intelligence: The state of the art*. IOS press, 2021, pp. 1–51.
- [32] A. Garcez, M. Gori, L. Lamb, L. Serafini, M. Spranger, and S. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," *Journal of Applied Logics*, vol. 6, no. 4, pp. 611–631, 2019.
- [33] T. Rocktäschel and S. Riedel, "End-to-end differentiable proving," *Advances in neural information processing systems*, vol. 30, 2017.
- [34] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, "Deepproblog: Neural probabilistic logic programming," *Advances in neural information processing systems*, vol. 31, 2018.
- [35] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," *Advances in neural information processing systems*, vol. 30, 2017.
- [36] A. Capitanelli and F. Mastrogiorganni, "A framework for neurosymbolic robot action planning using large language models," *Frontiers in Neuro-robotics*, vol. 18, 2023.
- [37] Y. Liang, N. Kumar, H. Tang, A. Weller, J. B. Tenenbaum, T. Silver, J. F. Henriques, and K. Ellis, "Visualpredicator: Learning abstract world models with neuro-symbolic predicates for robot planning," *arXiv preprint arXiv:2410.23156*, 2024.
- [38] R. Wang, B. He, and M. Pajic, "Neuro-symbolic deadlock resolution in multi-robot systems," in *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, 2025, pp. 1066–1077.
- [39] K. Namasivayam, H. Singh, V. Bindal, A. Tuli, V. Agrawal, R. Jain, P. Singla, and R. Paul, "Learning neuro-symbolic programs for language guided robot manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7973–7980.
- [40] T. Silver, A. Athalye, J. B. Tenenbaum, T. Lozano-Pérez, and L. P. Kaelbling, "Learning neuro-symbolic skills for bilevel planning," in *6th Annual Conference on Robot Learning*, 2023.
- [41] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, and D. Zhou, "Neural logic machines," in *Int. Conf. on Learning Representations (ICLR)*, 2019.
- [42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
- [43] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The world wide web conference*, 2019, pp. 2022–2032.
- [44] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of the web conference*, 2020, pp. 2704–2710.
- [45] A. Josang and R. Ismail, "The beta reputation system," in *15th bled electronic commerce conference*, vol. 5, 2002, pp. 2502–2511.
- [46] Webots, "http://www.cyberbotics.com," open-source Mobile Robot Simulation Software. [Online]. Available: http://www.cyberbotics.com