MIRAGE: BENCHMARKING INTENT IDENTIFIABILITY UNDER PARTIAL OBSERVATION THROUGH COUNTERFACTUAL MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Intent inference from partial observations underpins autonomous systems, multi-agent interaction, and situation awareness, yet existing formulations assume that an observed trajectory determines a unique latent intent. This assumption fails in settings where different intentions can generate similar behaviors: a movement toward a region may correspond to an attack, a reconnaissance pass, or a deliberate feint, and single-label evaluation cannot distinguish poor reasoning from intrinsically ambiguous evidence. We introduce MIRAGE, a counterfactual benchmark for intent identifiability. MIRAGE generates matched counterfactual trajectories by intervening on latent intent while keeping scenario conditions fixed. We formalise the counterfactual equivalence set (CES), the set of intents compatible with an observation prefix at each step, using an explicit observation distance D_{obs} and a calibrated indistinguishability threshold ϵ_o, released as a criterion-indexed family and reported under its strictest member. We further introduce ShenJi, an evidence-aware set-valued reasoner that maintains multiple hypotheses and contracts them as discriminative evidence emerges, with contraction enforced by a cardinality regulariser rather than observed post hoc. Across ten runs, 38.7% of trajectory steps remain ambiguous under CES (88.5% within the evaluation window, a lower bound). ShenJi preserves the commanded intent in 94.5% of ambiguous windows while reducing premature exclusion to 0.006, compared with 0.090 and 0.157 for deep-ensemble and conformal baselines. However, discrimination on resolved windows remains below chance level. Selective abstention and adaptive conformal prediction cannot recover the missing evidence, indicating that the observation contract, rather than estimator capacity alone, limits identifiability.

1 INTRODUCTION

Inferring the intent behind observed behavior is a central problem in artificial intelligence, underpinning autonomous systems, multi-agent interaction, and situation awareness. Existing intent recognition(Liu et al., 2021; Chen et al., 2023; Yang et al., 2022; Wang et al., 2023) typically assumes that partial observations determine a unique latent intent. However, under partial observation, multiple intentions may generate similar behaviors, making prediction errors indistinguishable from intrinsic ambiguity. Therefore, intent recognition requires measuring not only prediction accuracy but also whether the available evidence permits identification.

We introduce MIRAGE as a response to this gap: a counterfactual benchmark for intent identifiability under partial observation. As shown in Figure 1, MIRAGE generates matched counterfactual trajectories by controlling latent intent variables while keeping scenario conditions fixed, including environment, initial state, and controller configuration. This construction enables direct comparison between alternative intentions that produce similar observable behaviors and provides explicit ground truth for measuring ambiguity. Based on this formulation, we define the counterfactual equivalence set, the set of intents compatible with observations at each time step, shifting the objective from forced single-label prediction toward evidence-dependent set-valued reasoning. Because identifiability depends on what is considered indistinguishable, we explicitly define the observation distance and threshold used to measure ambiguity.

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.

To evaluate this objective, we introduce ShenJi, an evidence-aware set-valued reasoner that maintains a candidate intent set and contracts it as discriminative evidence accumulates. Its contraction is enforced through an explicit cardinality regulariser, making progressive narrowing a property of the objective rather than an observed side effect. Our experiments show that the observation contract, rather than the estimator alone, constrains intent identifiability.

Our contributions are summarized as follows:

- We establish *intent identifiability under partial observation* as a distinct problem, show that single-label evaluation conflates inference failure with identification ambiguity, and make the two separable by an identifiability-aware evaluation protocol.
- We build **MIRAGE**, a counterfactual benchmark providing synchronized latent intents, realized behaviors, and partial observations over 40 scenario–seed groups and three tasks, with an audit protocol—metadata-only baselines, input-channel auditing, and matched counterfactual pairs—showing that performance reflects behavioral reasoning rather than generation artifacts.
- We define the **counterfactual equivalence set** over an explicit observation distance D_{obs} and a calibrated indistinguishability threshold ε_o, released as a criterion-indexed family, and prove a monotone-contraction property (Proposition 1) that reports *when* an intent becomes identifiable, not only whether it is predicted.
- We show that identifiability is observation-limited rather than estimator-limited and introduce ShenJi, an evidence-aware set-valued baseline with explicit contraction regularization.

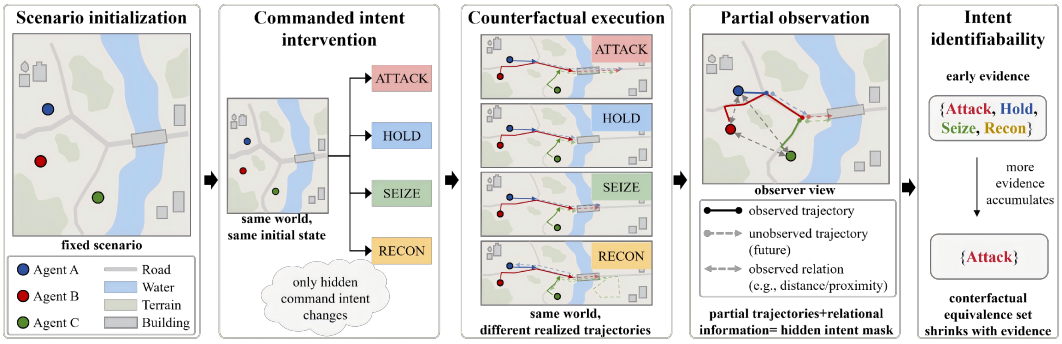


Figure 1: **MIRAGE separates intent, behavior, and evidence to make ambiguity measurable.** Counterfactual trajectories are generated under fixed scenario conditions while the commanded intent is varied, so differences between trajectories are attributable to intent rather than to environment or initial state. The observer sees only the partial observation channel, and the resulting counterfactual equivalence sets record which intents remain identifiable at each step

2 Related Work

Most intent recognition methods formulate intent as a single label inferred from a behavioral sequence and evaluate it using single-label accuracy. Existing intent-related studies in trajectory forecasting (Liu et al., 2021; Chen et al., 2023), tactical recognition (Yang et al., 2022; Wang et al., 2023; Cao et al., 2025; Song et al., 2026), dialogue (Zhang et al., 2024), and pedestrian behavior(Rasouli et al., 2019; Malla et al., 2020) mainly evaluate single-label prediction. Classical plan recognition (Kautz & Allen, 1986), Bayesian models (Charniak & Goldman, 1993; Bui et al., 2002), POMDPs (Kaelbling et al., 1998; Bui, 2003), and inverse reinforcement learning (Ramachandran & Amir, 2007; Ziebart et al., 2008) model latent goals but still assume a predefined inference target. Under partial observation, this protocol cannot distinguish ambiguity caused by the observation from errors introduced by the method; MIRAGE therefore evaluates the compatible-intent set itself.

Intent identifiability is closely related to uncertainty estimation in partially observable decision processes. Existing POMDP and Bayesian approaches represent uncertainty over latent states of a specified model, while active perception reduces uncertainty by selecting informative observations (Kaelbling et al., 1998; Bui, 2003; Charniak & Goldman, 1993; Bajcsy, 1988). Unlike active inference,

where observations can be selected to reduce uncertainty, MIRAGE assumes a fixed observation function and evaluates whether the resulting evidence is sufficient for identification. The observation contract is therefore the object being measured.

Existing trajectory benchmarks primarily focus on behavior prediction or policy learning. SMACv2 (Ellis et al., 2023) introduces partial observability but does not evaluate intent inference. Pedestrian and conversational datasets (Rasouli et al., 2019; Malla et al., 2020; Zhang et al., 2024) annotate observable behavioral intentions rather than hidden intents separated from their realizations. Causal and counterfactual benchmarks (Pearl, 2009; Khemakhem et al., 2020; Chalupka et al., 2015) study latent variable recovery but do not consider sequential trajectories with evolving ambiguity. None records the set of intents compatible with an observation prefix, which determines whether unique prediction is well posed. MIRAGE complements these benchmarks by evaluating which intents remain identifiable over time.

Set-valued prediction provides an alternative to forced single-label decisions under uncertainty. Conformal prediction provides coverage guarantees for prediction sets (Vovk et al., 2005; Angelopoulos & Bates, 2021), set-valued classification balances coverage and set size (Sadinle et al., 2019), and selective prediction enables abstention under uncertainty (Geifman & El-Yaniv, 2017; 2019). Partial-label learning considers static candidate-label ambiguity during training (Tian et al., 2023; Wang et al., 2025). These methods target model-relative uncertainty or a fixed candidate space, whereas CES changes with observation history as evidence accumulates. Ambiguity is therefore a dynamic property of the observation process rather than a static property of the label space. ShenJi follows this perspective; Section 6 evaluates conformal and selective alternatives under the same protocol.

3 Intent Identifiability under Partial Observation

Conventional intent recognition assumes that a partial observation sequence maps to a unique latent intent. Under partial observation, however, multiple intents may produce indistinguishable behaviors. We therefore formulate intent recognition as an identifiability problem.

Let a complete episode be $\tau = (s_1, \ldots, s_T)$, where s_t denotes the global system state, and each episode is associated with a latent intent $z \in \mathcal{Z}$. The partial observation at time t is defined as $o_{\leq t} = \mathcal{O}(s_1, \ldots, s_t)$, where $\mathcal{O}$ is the observation function defined by visibility constraints. Existing methods implicitly assume that the set of compatible intents contains only a single element, whereas in real-world scenarios, there are often cases involving more than one. The objective is therefore not always to predict a single intent, but to estimate the set of intents identifiable from the current evidence.

To quantify intent ambiguity, we introduce the counterfactual equivalence set (CES), which contains latent intents whose counterfactual observation prefixes remain indistinguishable from the reference prefix. The benchmark ground truth uses the strict identifiability limit to measure the earliest distinguishability boundary, while calibrated thresholds are provided for sensitivity analyses. Reproducibility requires three explicit choices: the definition of indistinguishability, measurement resolution, and tolerance, which are specified in Definitions 1–3. The CES is therefore a family indexed by $(D_{\text{obs}}, \epsilon_o)$ rather than a single object. Section 6.2 shows that different instantiations produce different ambiguity levels and that the released MIRAGE setting provides a lower bound on all reported ambiguity measurements.

Definition 1 (Observation distance)**.** *Let $O^a_{1:t}$ and $O^b_{1:t}$ be two observation prefixes generated by the same observation function $\mathcal{O}$. At step s, let E_s denote the entities visible in both prefixes, and let $\text{hex}(e, s)$, $v(e, s)$, and $\rho(e, s)$ denote the grid position, speed, and distance to the centroid of the visible entity set of e, respectively. The observation distance is the weighted mean of three normalized per-step discrepancies,*

$$D_{\text{obs}}\big(O^a_{1:t}, O^b_{1:t}\big) = \frac{1}{t} \sum_{s \leq t} \Big[\alpha\, d_{\text{pos}}(s) + \beta\, d_{\text{vel}}(s) + \gamma\, d_{\text{rel}}(s)\Big], \qquad \alpha + \beta + \gamma = 1, \tag{1}$$

where H is the map diameter and V is the speed range, ensuring that all components are normalized to $[0, 1]$. The three components are computed as the mean discrepancies over E_s for position, speed, and relational structure, respectively (Appendix C).

d_{pos} *measures trajectory divergence,* d_{vel} *measures observable kinematic divergence, and* d_{rel} *measures relational structure divergence, which is invariant to global translation.*

Definition 2 (Counterfactual equivalence set)**.** *Given a reference intent* $z^\star$ *and its observation prefix* $O^{z^\star}_{1:t}$*, the CES at time* t *is defined as*

$$\mathcal{C}^\star_t = \left\{ z \in \mathcal{Z} \;\middle|\; D_{\text{obs}}\left(O^z_{1:t}, O^{z^\star}_{1:t}\right) \le \varepsilon_o \right\}, \tag{2}$$

where $O^z_{1:t}$ *denotes the observation prefix obtained by executing intent* z *under identical scenario conditions. A singleton* $\mathcal{C}^\star_t$ *indicates an identifiable intent, whereas* $|\mathcal{C}^\star_t| > 1$ *indicates unresolved ambiguity. Since smaller* ε_o *imposes stricter equivalence, the limit* $\varepsilon_o \to 0^+$ *yields the smallest CES in the family; we use this ordering in Section 6.2.*

Definition 3 (Indistinguishability threshold)**.** *Let* $\mathcal{S}$ *denote a set of same-intent episode pairs, where each pair shares the same latent intent and scenario conditions but differs in stochastic execution factors (e.g., execution seed or controller implementation for the same commanded intent). The threshold is defined as the upper* q*-quantile of the same-intent distance distribution,*

$$\varepsilon_o = Q_q \left\{ D_{\text{obs}}(O_i, O_j) : (i, j) \in \mathcal{S} \right\}, \tag{3}$$

with $q = 0.95$*.*

Definition 3 makes the threshold an empirical property of the intent rather than a hand-tuned constant. Specifically, ε_o defines the boundary below which two executions of the same intent are considered indistinguishable, yielding a false-exclusion rate of $1 - q = 5\%$ by construction. Any pair below ε_o therefore represents a valid equivalence relation rather than numerical noise.

Proposition 1 (Monotone contraction)**.** *Fix a reference intent* $z^\star$ *and suppose the per-step discrepancy between* z *and* $z^\star$ *is eventually bounded away from zero:* $d(s) \ge \delta > \varepsilon_o$ *for all* $s \ge t_0$*. Then* $z \notin \mathcal{C}^\star_t$ *for every* $t \ge \varepsilon_o t_0/(\delta - \varepsilon_o)$*, so* $|\mathcal{C}^\star_t|$ *is non-increasing on that range and* $\mathcal{C}^\star_t \to \{z^\star\}$ *once every alternative intent has diverged.*

Ambiguity therefore resolves monotonically when different intents eventually diverge, with the resolution time determined by ε_o and divergence magnitude rather than heuristics (proof in Appendix C). The released MIRAGE ground truth uses the exact-signature limit of (1), with (α, γ, β) approaching one-hot weighting over observable fields and $\varepsilon_o \to 0^+$; Section 6.2 analyzes how calibrated variants affect the identifiability timeline.

4 MIRAGE: A Counterfactual Benchmark for Intent Identifiability

MIRAGE is a controlled benchmark designed to investigate whether latent intentions can be inferred from incomplete observations. Unlike conventional trajectory datasets, which typically assign a single intent label to each behavioral sequence, MIRAGE separates the underlying intent, the resulting behavior, and the available observational evidence. Rather than replacing existing intent recognition benchmarks, MIRAGE addresses a complementary question: whether an intent remains identifiable under controlled partial observations and counterfactual conditions, which is not explicitly evaluated by current datasets.

4.1 Counterfactual Benchmark Construction

The benchmark constructs counterfactual trajectories in which the underlying intent changes while irrelevant environmental factors remain fixed (Figure 1). An episode is generated as $\tau_z = \mathcal{G}(x_0, e, z, \pi, \omega)$ from an initial state x_0, environment configuration e, commanded intent z, execution policy π and stochasticity ω, and exposed to the observer through $O^z_{1:t} = H(\tau_z, M)$ with observation mask M. For a set of counterfactual executions $\mathcal{T}_{cf} = \{\tau_z \mid z \in \mathcal{Z}\}$ that share scenario context but differ in latent intent, MIRAGE derives the CES by comparing their observable consequences.

Generation proceeds in three stages. First, latent intent variables are assigned before execution, with each intent corresponding to a behavioral objective that influences subsequent agent decisions. Second, trajectories are generated under controlled scenario conditions, so that different intents may produce similar early-stage behaviors and ambiguity arises naturally. Third, observations are produced through partial observation functions: instead of complete simulator states, MIRAGE exposes only information available from the observer perspective. This creates cases where different latent intents induce observationally similar prefixes under the same observation contract, causing single-label intent recognition to become unreliable.

4.2 Observation Contract and Ground Truth

A central design objective is to prevent information leakage between intent labels and model inputs. For each timestamp the benchmark maintains three synchronized representations: the **global state** (complete simulator information, used only for ground truth), the **realized behavior** (actions and state transitions executed by agents), and the **partial observation** (information available to the observer). Different tasks use different subsets; an observation cutoff ensures that models cannot access future actions or post-event information. This separation allows MIRAGE to measure inference from incomplete evidence rather than recognition of generated labels. For each prefix the benchmark provides the CES $\mathcal{C}_t^\star$ of Definition 2, which starts large, contracts as discriminative evidence appears (Proposition 1). This allows evaluation of not only whether a model eventually predicts the right intent but whether it understands when an intent becomes identifiable.

4.3 Benchmark Tasks and Validation

MIRAGE supports three complementary tasks: **intent set prediction** ($\hat{C}_t = f(o_{\le t})$), **intent identifiability estimation** ($g(o_{\le t}) \to \mathbb{I}(|\mathcal{C}_t^\star| = 1)$), and **intent resolution forecasting** ($\Delta t = \min\{t : |\mathcal{C}_t^\star| = 1\}$), extending intent recognition from static classification to temporal reasoning.

Because synthetic benchmarks may be solved through generation artifacts rather than the intended reasoning capability, MIRAGE includes explicit validation protocols: metadata-only baselines over scenario and policy identifiers, cross-condition evaluation on unseen scenarios and policies, and an analysis of performance as a function of evidence level. Dataset statistics, the intent ontology, generation configurations, and the full validation results are provided in Appendix A.

5 ShenJi: Evidence-Aware Set-Valued Reasoning

MIRAGE is designed to evaluate intent identifiability rather than model accuracy alone. To study reasoning under ambiguity, we introduce ShenJi, an evidence-aware set-valued reasoner(Figure 2). Its objective is to maintain a candidate intent set and progressively remove unsupported hypotheses as evidence accumulates, rather than forcing early commitment to a single intent. ShenJi therefore requires evidence-aware scoring, adaptive thresholds, and explicit contraction regularization.

5.1 Evidence-Aware Intent Representation

Given an observation sequence $o_{\le t}$, ShenJi encodes temporal and relational evidence as $h_t = f_\theta(o_{\le t})$, representing the accumulated evidence state. A typed message-passing layer over the intent-relation graph produces per-intent representations, and ShenJi estimates intent compatibility $s_t(z) = g_\phi(h_t, e_z) \in [0, 1]$, where e_z is the intent embedding and $s_t(z)$ measures whether intent z is supported by current evidence. Unlike a single-label predictor, s_t is trained as a membership score against the CES, with set membership rather than exclusivity as the calibration target.

5.2 Evidence Insufficiency and the Adaptive Threshold

A fixed threshold faces a trade-off: a conservative value preserves valid hypotheses under ambiguity but retains unsupported candidates indefinitely, whereas an aggressive value produces smaller sets by prematurely excluding intents that later prove correct. Avoiding this trade-off requires the decision boundary to depend on the amount of missing evidence, which is explicitly modeled.

Definition 4 (Evidence insufficiency)**.** *The evidence-insufficiency head h_ψ maps the evidence state to a scalar $u_t = h_\psi(h_t) \in [0, 1]$, the degree to which the accumulated evidence fails to separate the*

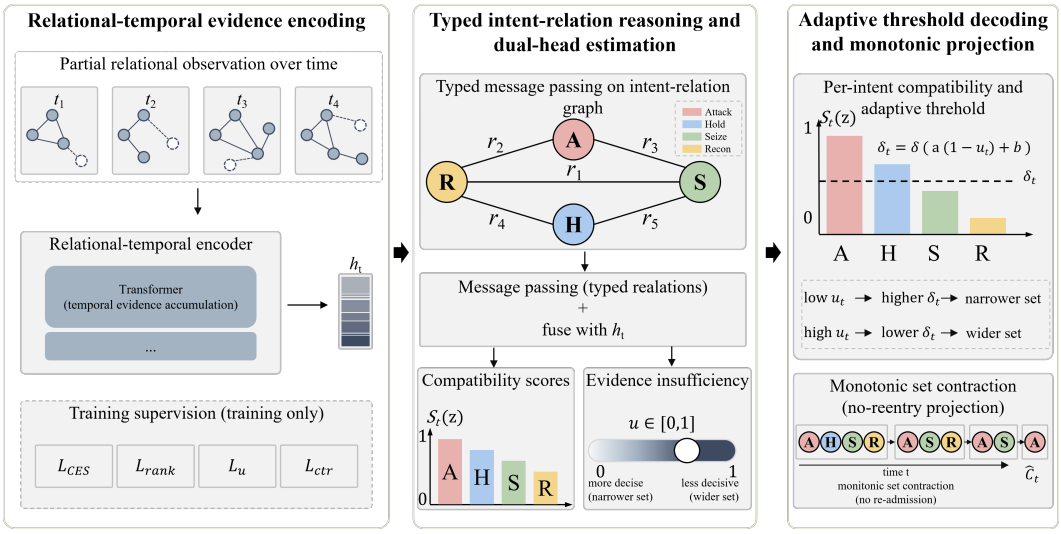


Figure 2: **ShenJi maintains and contracts a hypothesis set instead of committing to one intent.** The encoded evidence state is scored for compatibility with each intent, and an evidence-insufficiency estimate drives an adaptive decision threshold: scarce evidence widens the prediction set, while discriminative evidence contracts it.

currently compatible intents. It is supervised by the ground-truth identifiability level $u_t^\star = 1 - 1/|\mathcal{C}_t^\star|$*, which is* 0 *once the CES is a singleton and approaches* 1 *as more intents remain compatible.*

Definition 5 (Adaptive threshold and prediction set)**.** *With learned scalars* $a, b \in \mathbb{R}$*, the decision threshold and prediction set are*

$$\delta_t = \sigma(a\,(1 - u_t) + b)\,, \qquad \hat{C}_t = \{\, z : s_t(z) \geq \delta_t \,\}\,, \tag{4}$$

where σ *is the logistic function. Because* δ_t *is decreasing in* u_t*, high insufficiency lowers the threshold and widens the set, while discriminative evidence raises the threshold and contracts it. Equation* (4) *parameterises the decision boundary with two learned scalars rather than a hand-set margin, allowing the contraction schedule to be fitted from data.*

5.3 Contraction-Regularised Objective

Progressive narrowing is desirable but is not guaranteed by membership supervision alone. We therefore add an explicit cardinality penalty that forbids the prediction set from growing along the trajectory,

$$\mathcal{L}_{\text{ctr}} = \frac{1}{T}\sum_{t=1}^{T-1} \max\Big(0,\ |\hat{C}_{t+1}| - |\hat{C}_t|\Big)\,, \tag{5}$$

enforcing $|\hat{C}_{t+1}| \leq |\hat{C}_t|$ in expectation. Since $|\hat{C}_t|$ is a non-differentiable count we substitute the soft cardinality $|\hat{C}_t| = \sum_z \sigma((s_t(z) - \delta_t)/\tau)$ with $\tau = 0.05$, and use the hard count at inference. The remaining terms are the CES membership objective and a counterfactual ranking margin applied on $\{t : |\mathcal{C}_t^\star| = 1\}$, with $y_t(z) = \mathbb{I}[z \in \mathcal{C}_t^\star]$:

$$\mathcal{L}_{\text{CES}} = -\sum_z \big[y_t(z)\log s_t(z) + (1 - y_t(z))\log(1 - s_t(z))\big]\,, \qquad \mathcal{L}_{\text{rank}} = \Big[\gamma - s_t(z^\star) + \max_{z \neq z^\star} s_t(z)\Big]_+\,. \tag{6}$$

The full objective is $\mathcal{L} = \mathcal{L}_{\text{CES}} + \lambda_1\mathcal{L}_{\text{rank}} + \lambda_2\mathcal{L}_{\text{ctr}} + \lambda_u\mathcal{L}_u$, where $\mathcal{L}_u$ is the binary cross-entropy of u_t against $u_t^\star$; training and inference are summarised in Algorithm 1 (Appendix D).

Proposition 2 (Monotone contraction guarantee)**.** *Let* $\mathcal{L}_{\text{ctr}} = 0$ *at the optimum and suppose the compatibility ordering of intents does not change over* $[t_1, t_2]$*. If* u_t *is non-increasing on* $[t_1, t_2]$*, then* δ_t *is non-decreasing and* $\hat{C}_{t+1} \subseteq \hat{C}_t$ *for all* $t \in [t_1, t_2]$*; hence* $|\hat{C}_t|$ *is non-increasing. Without the ordering assumption, the no-reentry projection in Algorithm 1 restores nestedness at the output level.*

Proposition 2 specifies the conditions under which the prediction sets contract and defines the contraction property evaluated in Section 6.4. The complete architecture configuration, the training and inference procedure (Algorithm 1), and the proofs of Propositions 1 and 2 are provided in Appendix D.

6 Experiments and Benchmark Validation

Our experiments evaluate MIRAGE from five perspectives—ambiguity validity, behavioral reasoning, single-label failure modes, set-valued uncertainty resolution, and temporal contraction—starting from the observation criterion itself (Section 6.2), because the ambiguity measured by a benchmark depends on that criterion.

6.1 Experimental Setup

We compare ShenJi with representative sequence-based and graph-based baselines, including recurrent models, transformer-based models, and trajectory reasoning approaches. All methods receive identical observation inputs under the same observation cutoff protocol (Appendix E.1). Because the ground truth is set-valued, evaluation uses task-matched metrics(Appendix C),including: **Set-F1** (agreement with the ground-truth CES), **ASS** (average set size), **IC@1/IC@2** (ranking quality of the top predicted intents), **compatible intent retention**, **premature exclusion rate** (PER_{amb}: fraction of ambiguous cases where a valid intent is removed), and **contraction trajectory agreement** (CTR). All numbers are means over the ten-run matrix (5 split seeds $\times$ 2 training seeds), with $\pm$ the population standard deviation across runs.

6.2 Calibrating the Observation Distance

We first examine how much identifiability the observation contract itself permits: Definition 3 makes ε_o estimable, so we calibrate it on the benchmark itself rather than fixing it by hand: 392 same-intent pairs (identical commanded intent, same scenario, different seed) and 144 different-intent pairs (same scenario and seed, different command). With the balanced weighting the same-intent distance has mean 0.0066 (max 0.0232) against 0.0075 (max 0.0207) for different-intent pairs, giving $\varepsilon_o = Q_{95}(D_{\mathrm{obs}} \mid \text{same intent}) = 0.0185$. The false-exclusion rate on same-intent pairs is 5% by construction; the false-merge rate, the share of genuinely *different* intents still declared indistinguishable, is 95%. The overlap indicates that the observation contract limits the identifiability available to any estimator.

The criterion determines how much ambiguity is resolved. Under the exact-signature limit 236 of 288 pairs become distinguishable at median onset 0.30 of the trajectory; under a threshold calibrated on our held-out collection (0.0125) only 76 do; under the in-corpus threshold, 14 at median onset 0.93. The threshold is also population-specific, with the two calibrations differing by $1.5\times$. The released ground truth is therefore the strictest member of the family: it fixes the earliest distinguishability boundary and yields ambiguity shares that are lower bounds on the calibrated variants, while an analysis in which the ambiguity level itself is the quantity of interest can adopt the calibrated variant, released as a module with its calibration set. Full sensitivity grid in Appendix C.2.

6.3 Structured Ambiguity and Benchmark Validity

We next evaluate whether the benchmark contains genuine ambiguity or reduces to standard classification. As shown in Figure 3, the ground-truth counterfactual equivalence sets progressively contract as discriminative evidence accumulates, showing a temporal transition from ambiguous to identifiable intent states. Over the full trajectories, 38.7% of steps retain a counterfactual equivalence set of size ≥ 2 (Appendix B.3), arising from three mechanistically distinct sources—delayed expression (168 cases), observation masking (16: evidence exists but is unavailable to the observer), and behavioral non-expression (8: distinct intents produce near-identical behavior) Within the 250-step evaluation window the ambiguous share is higher (88.5%), because the window ends before most counterfactual pairs have separated; model comparison is therefore carried out deliberately in the evidence-limited regime. Both values are lower bounds because the released criterion is the least ambiguous CES variant.

We further audit whether performance reflects behavioral reasoning rather than simulation artifacts. Metadata-only baselines using scenario and policy identifiers remain at the majority-class floor (Appendix B.2), so contextual metadata carries no recoverable intent signal. An input-channel audit confirms that no observation feature reveals the generating intent or future behavior (Appendix B). Finally, under matched counterfactual pairs that share an identical observation prefix but differ in commanded intent, ShenJi maintains near-identical predictions (same-intent consistency 0.9999; different-intent ambiguity preservation 0.9880; Appendix B.4).

6.4 Single-label Comparison, Selective and Conformal Baselines

We then evaluate how single-label prediction behaves under conditions where no unique answer exists. Single-label methods optimize $\hat{z} = argmax_z p(z|o_{\leq t})$ and are forced to commit under ambiguous conditions where a unique answer does not exist. Figure 3(B,C) shows that ShenJi matches the evolving ambiguity structure by adapting prediction set size and reducing premature exclusion. ShenJi contracts only against accumulated evidence, achieves Set-F1 0.767±0.026 (set size 2.66) over the ten-run matrix, and reduces premature exclusion on ambiguous cases to 0.006 on the matched-counterfactual protocol, versus 0.090 (deep ensemble) and 0.157 (conformal/energy-based) under the identical protocol (Appendix B.5). For reference, the degenerate always-all and always-prior predictors attain Set-F1 0.855 and 0.409; neither contracts, and Set-F1 alone is maximized by a predictor that carries no evidence, so MIRAGE reads set size and contraction jointly.

The IC@1 value is 0.280±0.005, and we treat this number as a diagnostic rather than a standalone result: it sits below the group-conditional uniform chance level 0.319 and close to the majority-class floor 0.278, so on its own it can be explained equally well by intrinsic ambiguity, weak discrimination, or class imbalance. We therefore report IC@1 separately on the identifiable and ambiguous subsets (Table 1).

Table 1: **Decomposing accuracy by ground-truth ambiguity localizes the failure.** Metrics on the 198 evaluation windows, split by ground-truth CES cardinality ($n = 30$ singleton, $n = 168$ ambiguous); ten-run means. IC@1 is higher on ambiguous windows (0.293) than on resolved ones (0.187), where it falls below the group-conditional chance level of 0.319: the aggregate score is limited by post-resolution discrimination, not by ambiguity. Ambiguous windows achieve top-$|\mathcal{C}_t^\star|$ retention 0.945 and Set-F1 0.828, so set size is not inflated to game the metric.

Subset	n	IC@1	top-K ret.	Set ret.	Set-F1	ASS
Singleton ($\|\mathcal{C}_t^\star\| = 1$)	30	0.187±0.078	0.187	0.660	0.334	2.43
Ambiguous ($\|\mathcal{C}_t^\star\| > 1$)	168	0.293±0.006	0.945	0.841	0.828	2.69
All windows	198	0.280±0.005	0.853	0.819	0.767	2.66

Two findings follow. First, IC@1 on the ambiguous subset (0.293) is *higher* than on the singleton subset (0.187): the aggregate IC@1 is not depressed by ambiguity, it is depressed by the model's failure to discriminate *after* the evidence has resolved, where it falls below chance. Second, the model is not gaming the metric by emitting large sets: on ambiguous windows it places the commanded intent in the top-$|\mathcal{C}_t^\star|$ scores 0.945 of the time and inside the decoded set 0.841 of the time, with set-level agreement Set-F1 0.828. The results show that set-valued reasoning preserves ambiguity, while post-resolution discrimination remains challenging.

A residual error limited by the evidence rather than by confidence should resist both standard remedies, which we therefore evaluated. For selective prediction we use the score margin $s_{(1)} - s_{(2)}$ and trace the risk–coverage curve. A confidence signal informative about correctness should raise accuracy sharply as coverage falls; it does not. ShenJi attains IC@1 0.282 at 50% coverage against a no-abstention value of 0.280, with area under the risk–coverage curve 0.734, while a temperature-calibrated single-label reduction of the same score vector reaches only 0.723; the oracle selective predictor, which abstains exactly on the windows it gets wrong, reaches 0.560 at 50% coverage. The limited selective gain indicates an evidence-limited rather than confidence-limited error.

For conformal prediction we implement group-disjoint split conformal (leave-one-group-out over test groups) and adaptive conformal inference, which maintains an online score threshold $q_{t+1} = q_t + \eta(\alpha - \mathrm{err}_t)$ (Gibbs & Candès, 2021). At $\alpha = 0.1$ adaptive conformal attains coverage 0.893

with mean set size 2.82 and Set-F1 0.795, against ShenJi's retained-command rate of 0.819 at set size 2.66 and Set-F1 0.767. At matched coverage adaptive conformal is the more set-efficient predictor (coverage 0.810 at set size 2.48 versus ShenJi's 0.819 at 2.66), while ShenJi is better on the benchmark's own objective (Set-F1 0.767 versus 0.728 at that operating point) and is the only method whose sets are constrained to contract over time. The two readings answer different questions, and we report both; full curves are in Appendix F.

6.5 Human Baseline under Partial Observation

The benchmark's central hypothesis yields a testable prediction: if the limitation arises from the available evidence rather than the estimator, human observers should also be unable to commit. Two independent human annotators first labelled 250 replay segments from partial (blue-side) observations under a written protocol with an adjudication pass (Appendix G); they agree substantially ($\kappa = 0.736$, Jaccard 0.768) and do *not* commit, with abstention rates of 24.5% and 54.5%. To test the prediction on the exact windows used for model comparison, we added a matched study on the 154 windows the models are scored on, under the same four-command protocol and a pre-registered Krippendorff's $\alpha \geq 0.7$ gate. This round passes the gate with high reliability ($\alpha = 0.9713$, identical sets on 150/154), and the annotators again decline to commit: no window receives a singleton set, mean set size is 2.81 and 2.83, and no abstentions are recorded. They retain the commanded intent in 89.6% and 92.2% of windows ($\mathrm{PER}_{amb} = 2.3\%$), an operating point not yet reached by the learned model (Section 6.4 reports 0.819 retention at set size 2.66). The protocol requested set output only, so no human single-label prediction is available for an IC@1 comparison; the human data therefore constrain identifiability rather than discrimination. Both studies point to the same cause: the observation contract, not the estimator, limits identifiability.

6.6 Open-Set Intents and Temporal Resolution

A closed intent ontology may not hold in deployment settings, so we also evaluate whether a model trained on the closed command space can flag out-of-ontology behaviour. Treating RECON as novel relative to the $\{$ATTACK, HOLD, SEIZE$\}$ ontology, the AUROC of the natural novelty scores—maximum compatibility, decision margin, and negative evidence insufficiency—is at chance (0.52, 0.502, 0.460), and an abstention threshold at the 10th percentile of known-class confidence flags 15.5% of novel windows against 10.1% of known windows, too small a separation to be useful (Appendix H). The failure is related to representation rather than confidence calibration: closed-set training does not produce an UNKNOWN detector because the model lacks a representation for intents outside the ontology. This negative result isolates a second open problem alongside post-resolution discrimination.

Finally, a system that truly tracks identifiability should satisfy $|\hat{C}_{t_1}| \geq |\hat{C}_{t_2}|$ for $t_1 < t_2$ and contract where the ground-truth set contracts. ShenJi's sets exhibit this monotone contraction, with CTR 0.344 ± 0.254—a property that static set-construction methods do not model and that the cardinality regulariser of Section 5 enforces by construction (Appendix E).

7 Discussion and Conclusion

Existing benchmarks evaluate predefined intent labels under the assumption that observed behaviors uniquely determine intent. MIRAGE instead measures the set of intents supported by available evidence, distinguishing inference errors from intrinsic ambiguity. Our experiments show that set-valued reasoning preserves unresolved ambiguity, while discrimination after evidence accumulation remains challenging. This limitation is shared across models, selective prediction, conformal calibration, and human observers, suggesting that identifiability is constrained by the observation process rather than estimator capacity. Future work may extend MIRAGE to open-set intents, active evidence acquisition, and broader observation settings.

In conclusion, MIRAGE reformulates intent recognition under partial observation as an identifiability problem rather than forced classification. By providing matched counterfactual trajectories and criterion-indexed ground truth, MIRAGE enables systematic evaluation of ambiguity and evidence-driven intent resolution. Experiments demonstrate that set-valued reasoning effectively preserves ambiguity, while discrimination under limited evidence remains an open challenge.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Hung H. Bui. A general model for online probabilistic plan recognition. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 1309–1318. Morgan Kaufmann, 2003.

Hung H. Bui, Svetha Venkatesh, and Geoff West. Policy recognition in the abstract hidden markov model. *Journal of Artificial Intelligence Research*, 17:451–499, 2002.

Bo Cao, Qinghua Xing, Longyue Li, and Weijie Lin. An air target intention data extension and recognition model based on deep learning. *Scientific Reports*, 15(1):13894, 2025.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Eugene Charniak and Robert P. Goldman. A bayesian model of plan recognition. *Artificial Intelligence*, 64(1):53–79, 1993.

Li Chen, Xingxing Liang, Yanghe Feng, Longfei Zhang, Jing Yang, and Zhong Liu. Online intention recognition with incomplete information based on a weighted contrastive predictive coding model in wargame. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7515–7528, 2023.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1660–1672, 2021.

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

Henry A. Kautz and James F. Allen. Generalized plan recognition. In *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI)*, pp. 32–37, 1986.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Xuan Liu, Meijing Zhao, Song Dai, Qiyue Yin, and Wancheng Ni. Tactical intention recognition in wargame. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pp. 429–434, 2021.

Srikanth Malla, Behzad Dariush, and Chiho Choi. TITAN: Future forecast using action priors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11186–11196, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2586–2591, 2007.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6261–6270, 2019.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Yafei Song, Lemin Li, Wen Quan, Peng Ni, Wen Sun, Ke Wang, and TianPeng Li. An architecture for tactical intention recognition of aerial targets based on unsupervised momentum contrast and transformer. *Journal of Big Data*, 13(1):67, 2026.

Yingjie Tian, Xiaotong Yu, and Saiji Fu. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 161:708–734, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Siyuan Wang, Gang Wang, Qiang Fu, Yafei Song, Jiayi Liu, and Sheng He. STABC-IR: An air target intention recognition method based on bidirectional gated recurrent unit and conditional random field with space-time attention mechanism. *Chinese Journal of Aeronautics*, 36(3):316–334, 2023.

Wei Wang, Dong-Dong Wu, Jindong Wang, Gang Niu, Min-Ling Zhang, and Masashi Sugiyama. Realistic evaluation of deep partial-label learning algorithms. In *International Conference on Learning Representations*, 2025.

Zhen Yang, Zhi-xiao Sun, Hai-yin Piao, Ji-chuan Huang, De-yun Zhou, and Ren Zhang. Online hierarchical recognition method for target tactical intention in beyond-visual-range air combat. *Defence Technology*, 18(8):1349–1361, 2022.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations (ICLR)*, 2024.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

AI USE STATEMENT

Generative AI tools were used only for manuscript polishing and language refinement. All scientific content, including methodology, experiments, code, benchmark construction, results, numerical analyses, citations, and interpretations, was independently conducted, reviewed, and verified by the authors. The authors are solely responsible for the final content of this paper.

REPRODUCIBILITY STATEMENT

All experiments use the released MIRAGE benchmark (40 scenario–seed groups; commanded space $\mathcal{C} = \{\text{ATTACK}, \text{HOLD}, \text{SEIZE}, \text{RECON}\}$) under a multi-split protocol (5 split seeds $\times$ 2 training seeds; splits, per-run predictions, and the ten-run results matrix are in the supplementary material). The observation distance and threshold of Definitions 1–3 are released as a standalone module together with the calibration set used in Section 6.2; Architecture, hyperparameters, and the margin-threshold sweep are in Appendix D; metric definitions in Appendix C. No validation or test data were used for model selection beyond the stated protocol. Benchmark data, splits, predictions, and evaluation scripts are in the anonymized repository (`https://anonymous.4open.science/r/mirage-benchmark-538C`).

ETHICS STATEMENT

The benchmark trajectories are generated by scripted and learned controller policies in simulation; annotation was limited to post-hoc verification of machine-generated records by the authors and research assistants. No personally identifiable information is contained in the data, and the released benchmark contains only simulated military scenario trajectories for research on intent reasoning under ambiguity.

ACKNOWLEDGMENTS

Omitted for anonymous review.