Value Identifiable, Credit Not: A Credit-First Identifiability Frontier for Confounded Sequential Decision-Making

Teresa Zhang

Stanford University, Stanford, CA, USA
teresaz@stanford.edu

Abstract. Temporal credit assignment seeks to isolate the specific decision-time interventions responsible for downstream returns. In observational trajectories, however, agent actions and environmental outcomes are frequently entangled by unobserved confounders. Consequently, standard correlation-based credit signals—such as temporal difference (TD) errors, advantage estimates, or multi-agent Shapley values—can yield systematically misleading attributions, even when aggregate policy value predictions appear highly accurate. While proximal causal inference mitigates hidden confounding for value-centric objectives via bridge-function moment restrictions, we demonstrate that value identifiability does not inherently guarantee credit identifiability. In this paper, we formalize per-timestep temporal credit as a sequential interventional estimand constrained by an explicit continuation policy. By analyzing identifiability through a discrete proximal moment equation $M\theta = y$, we establish a rigorous functional criterion: a linear functional $f^\top\theta$ is identifiable if and only if f resides within the row space of the moment matrix, $f \in \mathrm{Range}(M^\top)$. We provide an explicit finite-dimensional proof separating value identifiability from temporal-credit identifiability, demonstrating an irreducible identification width for credit even in the limit of infinite data. Empirically, we validate this predicted shrinkage-versus-plateau signature by comparing finite-sample interval widths against the oracle population width. Furthermore, we introduce a temporal responsibility metric alongside a computable row-space residual to serve as an operational "credit identifiability risk" diagnostic. Finally, using a controlled, tabular SCM illustrative environment, we empirically show the fundamental structural vulnerabilities of representative multi-agent reinforcement learning (MARL) baselines, demonstrating they exhibit systematic, sign-level misattribution under latent confounding, whereas our proximal interval credit remains demonstrably robust.

Keywords: Causal Reinforcement Learning · Temporal Credit Assignment · Proximal Causal Inference · Multi-Agent Systems · Partial Identification.

1 Introduction

Credit assignment poses a fundamental causal challenge in the deployment of reinforcement learning (RL) and multi-agent systems (MAS): determining precisely which historical decisions caused subsequent rewards. In sequential observational data, actions are rarely randomized; hidden environmental factors frequently confound the relationship between agent behavior and resulting outcomes. In these confounded settings, standard credit signals extracted from logged trajectories—such as Monte Carlo advantages, TD-errors, or observational counterfactual heuristics—can be systematically biased. Crucially, these attribution errors persist even when aggregate value metrics, such as expected return, appear benign (as we will empirically demonstrate in Section 8, Figure 3).

This paper establishes a credit-first framework for proximal identification in confounded sequential decision-making. When hidden confounding is addressed through proxy-based moment restrictions, the identifiable object is not the underlying "bridge" parameter θ itself, but rather particular linear functionals of θ. We demonstrate that while some functionals corresponding to policy value are identifiable, other functionals corresponding to temporal credit may remain structurally non-identifiable under the exact same observational distribution. This distinction is structural, not an artifact of finite-sample estimation. When credit is non-identifiable, any algorithm forcing a point-valued credit report necessarily encodes unverified assumptions beyond the maintained moment restrictions.

Implications for Agentic Systems and MAS. Autonomous agents routinely depend on credit assignment signals for critical operational tasks, examples including behavioral debugging, policy adaptation, blame allocation, and establishing trust in inter-agent communication. If the underlying causal credit is structurally ambiguous due to unmeasured confounding—yet the algorithmic architecture forces a crisp attribution—the resulting estimate embeds invisible biases into the agent's decision loop. This structural vulnerability directly aligns with the objectives of causal learning in multi-agent environments: causal reasoning must advance beyond merely predicting aggregate value to rigorously defining the boundary conditions under which causal explanations and local attributions are mathematically well-posed.

Contributions. We introduce a fundamentally credit-first framework for evaluating identifiability in sequential environments, structured as follows:

- **Interventional Credit Formalization:** We redefine per-timestep credit as an explicit interventional estimand, strictly coupling the action contrast with a defined continuation policy (Section 3).
- **Functional Identifiability Criterion:** For discrete proximal moment systems defined by $M\theta = y$, we establish that any linear functional $f^\top\theta$ is identifiable if and only if $f \in \mathrm{Row}(M)$ (Section 5).
- **Value-Credit Separation Proof:** We construct a rigorous finite-dimensional witness proving that temporal credit can remain structurally non-identifiable even when policy value is perfectly identified, yielding an irreducible identification width (Section 5).

- **Diagnostic Residuals and Evaluation:** We derive a row-space residual diagnostic, $\rho(c; \widehat{M})$, to quantify credit identifiability risk and evaluate methods through rigorous temporal responsibility recovery metrics (Sections 5 and 7).
- **Multi-Agent Confounding Analysis:** Through controlled tabular environments, we illustrate the fundamental structural vulnerabilities of representative MARL estimators under latent confounding, demonstrating systematic sign-level biases, and introduce proximal interval semantics that maintain attribution integrity (Section 8).

2 Related Work and Positioning

Our formulation isolates the credit-first identification target, distinguishing our approach from existing literature focused predominantly on policy evaluation or model-free value estimation.

Proximal Causal Inference and Bridge Moments. Proximal causal inference provides a mechanism to identify causal effects under unmeasured confounding by leveraging proxy variables (negative controls) to satisfy conditional moment restrictions [3,4,5]. In the context of sequential decision-making, proximal methods have been successfully adapted for off-policy evaluation (OPE). For instance, Bennett and Kallus [6] introduced proximal reinforcement learning to achieve efficient OPE in partially observed Markov decision processes (POMDPs). However, these approaches uniformly target policy value. Our work diverges by shifting the identification target specifically to temporal credit, proving that the success of value identification does not guarantee the success of credit identification.

Credit Assignment in RL and MAS. Conventional multi-agent credit assignment techniques emphasize variance reduction and learning stability. Counterfactual baselines like COMA [7] and Shapley-value formulations [8,9,11] rely on observational conditional distributions to isolate individual agent contributions. While effective in fully observable settings, these methods degrade under confounding. Recently, Wang et al. [15] introduced MACCA, utilizing Dynamic Bayesian Networks (DBNs) to achieve causal credit assignment in offline MARL. While MACCA assumes strict DBN structure to achieve point identification, our proxy setting relies on fundamentally different structural assumptions; directly implementing MACCA within our proxy-based SCMs is therefore outside our scope. We audit the fundamental correctness of attribution under hidden confounding, demonstrating that without proper proximal adjustment, current MARL baselines converge to precise but causally incorrect point estimates under latent confounding.

Partial Identification and Set-Valued Semantics. When maintained assumptions fail to point-identify a causal parameter, standard practice dictates the use of partial identification and set-valued semantics [12,14]. We adapt this paradigm to sequential reinforcement learning by formalizing the "identification width" of a temporal credit functional bounded by proximal moment restrictions, allowing agents to explicitly compute and report attribution ambiguity.

3 Problem Formulation

3.1 Sequential Setting and Confounding

We examine episodic sequential decision processes of horizon T with a discount factor $\gamma \in [0, 1]$. At each timestep t, an agent observes a history $H_t = (O_1, A_1, \ldots, O_t)$, executes an action $A_t \in \mathcal{A}$, receives a scalar reward R_t, and transitions to the next state. The trajectory data is generated by a behavior policy π^b operating within an observational environment where unobserved latent variables U_t confound the causal relationship between actions and subsequent rewards. Let $G_t = \sum_{j=t}^{T} \gamma^{j-t} R_j$ denote the return-to-go.

To address the hidden confounder U_t, we require standard proximal independence conditions to hold sequentially.

Assumption 1 (Sequential Proximal Proxies). *For each timestep t, there exist observed proxies Z_t (treatment-inducing) and W_t (outcome-inducing) and an unobserved confounder U_t such that, conditional on the discretized history $\phi(H_t)$:*

1. $Z_t \perp (R_{t:T}, W_t) \mid (A_t, U_t, \phi(H_t))$ *(negative control for outcomes)*
2. $W_t \perp A_t \mid (Z_t, U_t, \phi(H_t))$ *(negative control for treatments)*

We assume sufficient overlap to estimate the necessary conditional distributions without degenerate cells.

To ensure tractability in the finite-dimensional linear system, the mapping function $\phi(H_t)$ discretizes the continuous history space into a finite tabular regime. In our tabular SCM experiments, U_t is sampled i.i.d. across time and proxy emissions depend only on the current U_t, guaranteeing Assumption 1 holds strictly for future returns. The theoretical analysis abstractly models these dynamics through general moment restrictions.

3.2 Credit as a Sequential Interventional Estimand

Standard reinforcement learning metrics assess expected observational returns. However, assigning per-timestep credit requires an inherently interventional lens: it isolates the specific causal impact of altering a single action at time t while strictly controlling the behavior of the agent in all subsequent timesteps.

Definition 1 (Interventional Credit with Continuation Policy). *Let π^{cont} denote a fixed continuation policy dictating agent behavior for all timesteps $t+1 \ldots T$ following an intervention. We define the conditional interventional Q-functional as:*

$$Q_t^{\text{do}}(h, a; \pi^{\text{cont}}) = \mathbb{E}\left[G_t \mid H_t = h, \text{do}(A_t = a), A_{t+1:T} \sim \pi^{\text{cont}}\right]. \tag{1}$$

For a given history h and a target action contrast between a and a', the formal temporal credit is defined as:

$$\tau_t(h; a, a'; \pi^{\text{cont}}) = Q_t^{\text{do}}(h, a; \pi^{\text{cont}}) - Q_t^{\text{do}}(h, a'; \pi^{\text{cont}}). \tag{2}$$

Crucially, the continuation policy π^{cont} is not a nuisance parameter; it is a fundamental constituent of the credit estimand. Depending on the agent's objective, π^{cont} may represent the historical behavior policy for retrospective debugging or an optimal target policy for prospective planning. In our synthetic evaluations, π^{cont} is explicitly locked within the simulator to establish a ground-truth benchmark for causal attribution.

3.3 A Schematic Causal Model with Proxies (Conceptual)

Figure 1 provides a schematic view: an unobserved confounder influences both actions and outcomes, and observed proxies carry information about the confounder. This figure is illustrative; our formal analysis proceeds through the induced moment restrictions.

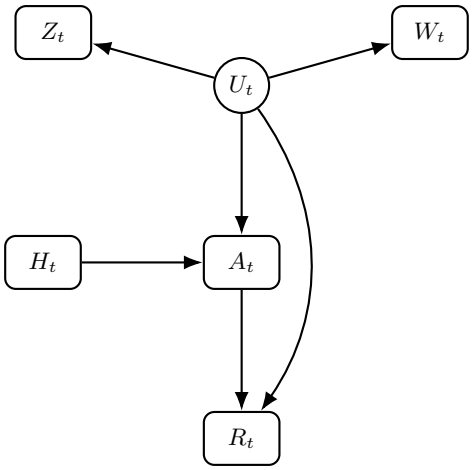

Fig. 1: Schematic confounded decision setting with observed proxies.

4 From Proximal Moments to a Linear System

Proximal causal identification operates through conditional moment restrictions involving the observed proxies and an unknown bridge function. In discrete or properly discretized continuous settings, these restrictions reduce to a finite-dimensional linear system.

4.1 Sequential Stacking and Moment Matrices

In sequential environments, the single-step bridge function restriction must be extended to encompass the return-to-go G_t. For a discrete setting, the sequential

bridge moment restriction requires finding a bridge function $\theta_t(w, a, h)$ such that for all observed states of (t, h, z, a):

$$\begin{aligned}\mu_t(z, a, h) &:= \mathbb{E}[G_t \mid Z_t = z, A_t = a, \phi(H_t) = h] \\ &= \sum_w \mathbb{P}(W_t = w \mid Z_t = z, A_t = a, \phi(H_t) = h)\, \theta_t(w, a, h).\end{aligned} \tag{3}$$

By stacking these equations across all timesteps t, history configurations h, and proxy realizations z, a, we construct a comprehensive finite-dimensional linear system. Let $\theta \in \mathbb{R}^d$ represent the flattened vector of unknown bridge parameters across all discretized cells. We define the moment matrix $M \in \mathbb{R}^{m \times d}$ containing the relevant conditional proxy probabilities, and the outcome vector $y \in \mathbb{R}^m$ containing the conditional expected returns. The population-level proximal restriction is thus expressed concisely as:

$$M\theta = y \tag{4}$$

where the rows of M index (t, h, z, a) and the columns index (t, h, w, a). Because the bridge function may not be uniquely identified by the proxies (due to rank deficiencies in M), the set of mathematically valid solutions forms an affine subspace, defined as $\mathcal{S}(M, y) = \{\theta : M\theta = y\}$.

Lemma 1 (Credit is a Linear Functional). *Under the bridge moment restriction (Eq. 3) and a fixed continuation policy π^{cont}, the interventional credit $\tau_t(h; a, a')$ can be mathematically expressed as a linear functional $c^\top \theta$, where the sparse weight vector $c = f_{t,h,a} - f_{t,h,a'}$ is explicitly determined by the observable proxy distributions.*

Proof (Main Steps). Under sequential proximal identification, the interventional return can be identified by reweighting the bridge function according to the identifiable observable transition dynamics under the continuation policy π^{cont}. Specifically, $Q_t^{\text{do}}(h, a; \pi^{\text{cont}}) = \sum_w \alpha_{t,h,a}(w)\theta_t(w, a, h)$, where α is a known weight vector derived strictly from observational data under π^{cont}. By defining $f_{t,h,a}$ as the flattened vector of these explicit weights over the discretized space, we obtain $Q_t^{\text{do}} = f_{t,h,a}^\top \theta$. The contrast credit is therefore the direct linear functional $\tau_t = (f_{t,h,a} - f_{t,h,a'})^\top \theta = c^\top \theta$.

4.2 Bounded Bridge Class and Feasibility

To guarantee that the partial identification width of non-identifiable linear functionals remains finite, we must restrict the parameter space.

Assumption 2. *There exists a finite constant $B > 0$ such that the true underlying bridge parameter satisfies $\|\theta\|_\infty \leq B$. Defining the bounded class $\Theta(B) = \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq B\}$, we assume the feasible intersection is non-empty: $\mathcal{S}(M, y) \cap \Theta(B) \neq \emptyset$.*

We emphasize that Assumption 2 does not drive the fundamental theoretical separation between value and credit. We treat B as a user-specified sensitivity bound restricting the admissible bridge class; the resulting partial identification intervals scale with B. Its theoretical purpose is strictly to provide the necessary finite geometry to compute actionable interval semantics for structurally ambiguous functionals.

5 Identifiability Analysis: Value Can Be Identifiable While Credit Is Not

5.1 A Functional Identifiability Criterion

We first establish the geometric conditions under which any linear functional of the bridge parameter can be uniquely determined from the observational moments.

Theorem 1. *Consider the linear moment equation $M\theta = y$ with unknown $\theta \in \mathbb{R}^d$ and known (M, y). Let $\ell_f(\theta) = f^\top \theta$ be any linear functional (representing value, credit, or an arbitrary query). Then ℓ_f is point-identifiable from (M, y) if and only if $f \in \mathrm{Range}(M^\top)$, equivalently $f \in \mathrm{Row}(M)$.*

When a functional falls outside the row space, its value is not uniquely determined. To quantify this ambiguity, we define the identification width over the bounded class.

Definition 2 (Identification width). *Fix M, y, and the bounded class $\Theta(B)$. For a linear functional $\ell_f(\theta) = f^\top \theta$, its identification width is:*

$$\mathrm{Width}_f(M, y; B) = \sup_{\theta \in \mathcal{S}(M,y) \cap \Theta(B)} f^\top \theta - \inf_{\theta \in \mathcal{S}(M,y) \cap \Theta(B)} f^\top \theta. \tag{5}$$

5.2 Value-Credit Separation Witness

The core theoretical result of this paper demonstrates that policy value estimation and temporal credit assignment possess entirely distinct identifiability frontiers. Using a minimal discrete SCM, we present a finite-dimensional witness where a value functional is perfectly identifiable, yet a credit-relevant functional over the identical moment system is structurally ambiguous.

Theorem 2 (Value Identifiable, Credit Non-Identifiable). *Consider a decision system with binary variables $Z, A, W \in \{0, 1\}$. Define the bridge parameter vector $\theta = (\theta_{0,0}, \theta_{1,0}, \theta_{0,1}, \theta_{1,1})^\top$ indexed by (W, A). We construct the 4×4 moment system $M\theta = y$ where the rows correspond to (Z, A) states and columns correspond to (W, A) states:*

$$M = \begin{pmatrix} p^0_{0,0} & p^1_{0,0} & 0 & 0 \\ 0 & 0 & p^0_{0,1} & p^1_{0,1} \\ p^0_{1,0} & p^1_{1,0} & 0 & 0 \\ 0 & 0 & p^0_{1,1} & p^1_{1,1} \end{pmatrix}, \quad y = \begin{pmatrix} \mu_{0,0} \\ \mu_{0,1} \\ \mu_{1,0} \\ \mu_{1,1} \end{pmatrix} \tag{6}$$

where $p^w_{z,a} = \mathbb{P}(W = w \mid Z = z, A = a)$ *and* $\mu_{z,a} = \mathbb{E}[R \mid Z = z, A = a]$*. There exist valid joint distributions yielding conditional probabilities* $\{p^w_{z,a}\}$ *such that a value functional* v *satisfies* $v \in \mathrm{Row}(M)$*, ensuring* $v^\top\theta$ *is point-identifiable. Simultaneously, a linear attribution functional* c *(acting as a proxy for credit queries) satisfies* $c \notin \mathrm{Row}(M)$*, ensuring* $c^\top\theta$ *possesses an irreducible identification width strictly greater than zero.*

Numeric Witness. To ground Theorem 2, let $p^0_{0,0} = 0.5$, $p^1_{0,0} = 0.5$, $p^0_{0,1} = 0.25$, $p^1_{0,1} = 0.75$, $p^0_{1,0} = 0.5$, $p^1_{1,0} = 0.5$, $p^0_{1,1} = 0.75$, $p^1_{1,1} = 0.25$, and let $y = (0.5, 0, 0.5, 0)^\top$. Solving $M\theta = y$ yields the constraint $\theta_{0,0} + \theta_{1,0} = 1$ alongside exactly $(\theta_{0,1}, \theta_{1,1}) = (0, 0)$. The solution space is thus invariant along the null-space vector $\delta = (1, -1, 0, 0)^\top$. If we query a value functional $v = (1, 1, 0, 0)^\top$, we find $v^\top\delta = 0$; hence v is orthogonal to the null space ($v \in \mathrm{Row}(M)$) and is point-identified as $v^\top\theta = 1$. However, for a contrast functional representing a specific credit proxy query $c = (1, -1, 0, 0)^\top$, we find $c \notin \mathrm{Row}(M)$ because $c^\top\delta = 2 \neq 0$. The value of $c^\top\theta$ varies over an interval of width at least $\Omega(B)$, demonstrating irreducible identification width and meaning point-valued reporting of this attribution functional is impossible under the maintained moment restrictions. Because Lemma 1 demonstrates that formal temporal credit τ operates as a structurally analogous linear contrast, there exist environments where τ inherently suffers from this exact non-identifiability.

5.3 A Diagnostic Residual for Identifiability Risk

Because non-identifiability is a structural property of the proxy overlap, we introduce a computable diagnostic. For any candidate credit functional c, we define the normalized row-space residual:

$$\rho(c; M) = \frac{\|c - \Pi_{\mathrm{Row}(M)}(c)\|_2}{\|c\|_2} \tag{7}$$

where $\Pi_{\mathrm{Row}(M)}$ denotes orthogonal projection onto the row space of M. At the population level, Theorem 1 guarantees that $\rho(c; M) = 0$ if and only if credit is identifiable. In finite samples, where sample noise may cause $\widehat{M}$ to artificially achieve full rank, we compute $\rho(c; \widehat{M})$ using a truncated SVD row space with threshold τ_{svd} (see Appendix B for details). This functions as an operational early-warning signal, allowing system designers to quantify the precise "identifiability risk" of relying on a specific temporal credit assignment.

Corollary 1 (Unavoidable worst-case error of point credit reports). *Fix* M, y*, and a credit functional* c*, and consider the induced interval* $\mathcal{I} = \left[\inf_{\theta \in \mathcal{S}(M,y) \cap \Theta(B)} c^\top\theta, \sup_{\theta \in \mathcal{S}(M,y) \cap \Theta(B)} c^\top\theta\right]$*. For any procedure that outputs a point estimate* $\widehat{\tau}$ *measurable with respect to observational information summarized by* (M, y)*, the worst-case error is bounded below by half the identification width:* $\sup_{\tau \in \mathcal{I}} |\widehat{\tau} - \tau| \geq \frac{\mathrm{Width}_c(M,y;B)}{2}$.

6 Interval Semantics and Estimators

Given an estimated system $(\widehat{M}, \widehat{y})$, define a conservative feasible set

$$\widehat{\mathcal{S}} = \left\{ \theta \in \Theta(B) : \|\widehat{M}\theta - \widehat{y}\|_\infty \leq \varepsilon \right\}$$

with ε chosen to upper bound finite-sample moment error (Appendix B). For a credit functional c, the induced interval is

$$\left[\widehat{\tau}^L, \widehat{\tau}^U\right] = \left[\min_{\theta \in \widehat{\mathcal{S}}} c^\top \theta, \max_{\theta \in \widehat{\mathcal{S}}} c^\top \theta\right],$$

computed by two linear programs.

We emphasize that the computed partial identification interval width is a function of the domain-specific sensitivity bound B. While B does not govern the binary condition of structural identifiability (Theorem 1), it entirely dictates the operational utility of the interval semantics. In theoretically unbounded reward settings (where $B \to \infty$), the partial identification width approaches infinity, rendering the interval uninformative.

Practical reporting rule (what an agent should do). Given a credit query c (timestep, action contrast, continuation policy), estimate $(\widehat{M}, \widehat{y})$, compute $\rho(c; \widehat{M})$, and: (i) if ρ is near zero and the interval is narrow, report a point (with the interval); (ii) if ρ is not near zero or the interval is wide, report the interval and/or abstain ("credit not identified under maintained assumptions").

7 Experiments

To rigorously validate the theoretical separation between value and credit, and to benchmark the failure modes of standard MARL attribution methods, we developed a suite of discrete, tabular Structural Causal Models (SCMs). In these environments, the ground-truth interventional credit τ is computationally accessible via exact simulator rollouts, providing an absolute oracle against which algorithms are evaluated. Unless stated otherwise, we evaluate the identification width using a sensitivity bound of $B = 1.0$, chosen as a valid upper bound given maximum rewards in $[0, 1]$ and $\gamma = 0.95$.

7.1 Synthetic SCM Design and Regimes

The generative process for the sequential trajectories is governed by the following core components over discrete horizons $T \in \{3, 5\}$:

- **Confounder process:** $U_t \in \{0, 1\} \sim \text{Bernoulli}(0.5)$ sampled i.i.d. at each timestep.
- **Behavior policy:** $A_t \in \{0, 1\}$ where $\mathbb{P}(A_t = 1 \mid U_t) = \eta_{U_t}$. Confounding strength is modulated by η.

- **Reward injection:** A single fixed responsible timestep $t^\star$ is designated per environment (e.g., $t^\star = 2$ for $T = 3$). At this active timestep, the reward is a Bernoulli draw $R_{t^\star} \sim \text{Bernoulli}(P(R \mid A_{t^\star}, U_{t^\star}))$. At all other inactive timesteps $t \neq t^\star$, the reward is a background emission $R_t \sim \text{Bernoulli}(P(R \mid U_t))$.
- **Proxy emissions:** Valid proxies $Z_t, W_t \in \{0, 1\}$ are emitted conditionally on U_t. Because U_t is i.i.d. across time and proxies depend solely on the contemporaneous U_t, Assumption 1 strictly holds for all future returns.
- **Continuation policy:** The interventional target queries assume $\pi^{\text{cont}} = \pi^b$ for retrospective debugging analysis.

By modulating the conditional emission probabilities of the proxies, we instantiate two primary SCM regimes:

1. **Identifiable Regime:** The transition dynamics are configured such that the desired temporal credit functional c satisfies $c \in \text{Row}(M)$. The oracle identification width is mathematically guaranteed to be 0.
2. **Separation Regime (Value-Identifiable / Credit-Non-Identifiable):** The proxy overlap is weakened, inducing a rank-deficient moment matrix where policy value remains in the row space, but the credit functional $c \notin \text{Row}(M)$. Here, the oracle width is strictly positive, representing irreducible structural ambiguity.

To evaluate multi-agent settings, the SCM is expanded into a cooperative Dec-POMDP configuration featuring two and three agents. A latent environmental variable concurrently influences the localized action tendencies of all agents alongside the global team reward, introducing intense spatial and temporal confounding.

7.2 Metrics and Estimation

We estimate the empirical moment matrix $\widehat{M}$ and vector $\widehat{y}$ using conditional frequencies and means derived from n offline trajectories. As defined in Section 4, the rows of $\widehat{M}$ index (t, h, z, a) and columns index (t, h, w, a) after discretization. The specific credit query c is the sparse vector encoding $\tau_t(h; a, a')$ under π^{cont}. We evaluate the realized-trajectory credit vector $(\hat{\tau}_1, \ldots, \hat{\tau}_T)$ for each episode.

To isolate a single responsible timestep $\hat{t}$ from the predicted interval sequence $[\hat{\tau}_t^L, \hat{\tau}_t^U]$, the *worst-case selector* computes a robust magnitude score $s_t = \max(|\hat{\tau}_t^L|, |\hat{\tau}_t^U|)$, while the *midpoint selector* computes $s_t = (\hat{\tau}_t^L + \hat{\tau}_t^U)/2$. The predicted timestep is then $\hat{t} = \arg\max_t s_t$.

Performance is assessed via **Top-1 Temporal Responsibility Accuracy** (the frequency with which $\hat{t} = t^\star$). While sharing conceptual motivation with the Halpern-Pearl degree of responsibility [16], our metric explicitly evaluates sequential action selection. We evaluate **Kendall τ Rank Correlation** between true and estimated credit vectors. Finally, to capture systematic sign-level misattribution, we evaluate the **Sign Error Rate**: $\frac{1}{N}\sum_i \mathbf{1}[\text{sign}(\hat{\tau}_i) \neq \text{sign}(\tau_i)]$, which tracks how frequently an algorithm actively penalizes a beneficial action.

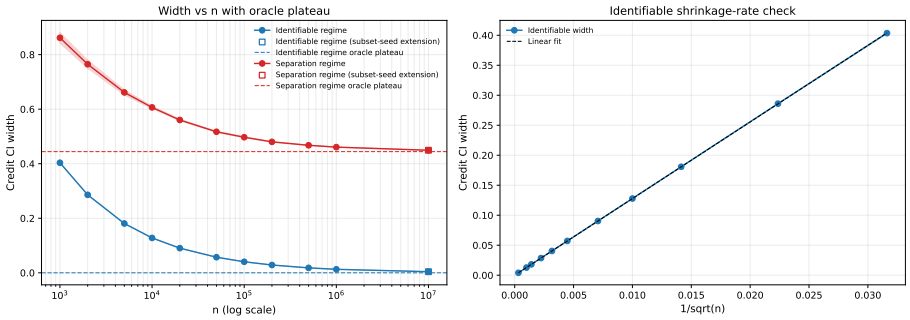

Fig. 2: **Shrinkage vs plateau in credit intervals.** Left: credit interval width versus sample size for identifiable and separation regimes, with dashed lines showing oracle (population) identification width. Square markers extend to $n = 10^7$ for a subset of seeds. Right: identifiable regime width versus $1/\sqrt{n}$ with a linear fit.

We baseline against Monte Carlo advantages (MC-Adv), TD-critic advantages (TD-Adv), observational differences (ObsDiff), and in multi-agent environments, COMA [7] and Shapley-style counterfactual credits [8,9,11].

8 Empirical Results

8.1 The Shrinkage vs. Plateau Signature

The empirical results prominently confirm the fundamental predictions of the functional identifiability criterion (Figure 2). In the identifiable regime, as the sample size n increases up to 10^7 trajectories, the width of the proximal credit interval systematically shrinks toward the population oracle width of 0. A regression analysis confirms this convergence strictly follows a $\mathcal{O}(1/\sqrt{n})$ rate, indicative of standard statistical estimation. Conversely, in the separation regime, increasing data volume yields diminishing returns. The finite-sample interval width initially contracts but abruptly plateaus at 0.4608 (at $n = 10^6$), nearly perfectly matching the theoretically derived oracle width of 0.4440 (see Table 1). This plateau is highly consistent with the theoretical prediction that credit ambiguity in confounded environments is a structural limitation of the proxy distribution, not an artifact of insufficient sampling or estimation noise.

8.2 Value-Credit Separation Under Confounding

To isolate the danger of relying on value metrics as a proxy for attribution quality, we performed a parameter sweep over the confounding strength parameter in the SCM (Figure 3). As confounding increases from 0.4 to 0.95, the estimation error for aggregate policy value remains uniformly low and stable, generally bounded

Table 1: Oracle (population) identification width and finite-sample interval width at $n = 10^6$ for Figure 2 (20 seeds).

Regime	Oracle id. width $\mathrm{Width}_c(M, y; B)$	Interval width @ $n = 10^6$
Identifiable	0.0000	0.0128
Value-id / Credit-nonid	0.4440	0.4608

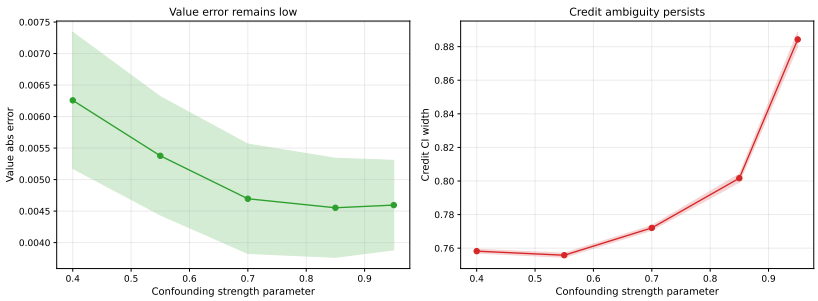

Fig. 3: **Value error vs. credit ambiguity under increasing confounding.** In separation regime, value error remains small while credit identification width (and its conservative interval width) remains large.

above by 0.0075. Simultaneously, however, the uncertainty (interval width) for the temporal credit functional inflates dramatically, rising above 0.88. This stark divergence validates the core premise of the paper: standard OPE diagnostics that report low value estimation error provide false confidence regarding the causal correctness of the underlying decision attributions.

8.3 Temporal Responsibility Recovery ($T = 3$ and $T = 5$)

When auditing timestep attribution across horizons $T = 3$ and $T = 5$ (Figures 4 and 5), correlation-based baselines—including Monte Carlo advantages, TD-critic advantages, and observational heuristics—illustrate severe vulnerabilities. Because these baselines condition on observational history rather than intervening on the confounder, they systematically absorb spurious correlations. Consequently, their Top-1 responsibility accuracy collapses to near 0.0, indicating they reliably select the *incorrect* timestep. In stark contrast, the proximal midpoint and worst-case interval selectors consistently identify the true responsible timestep, maintaining near-perfect Kendall τ alignment (approaching 1.0).

8.4 Residual Diagnostics and Stress Testing

Figure 6 demonstrates the operational utility of the row-space residual $\rho(c; \widehat{M})$. There is a strong positive association between the magnitude of the empirical

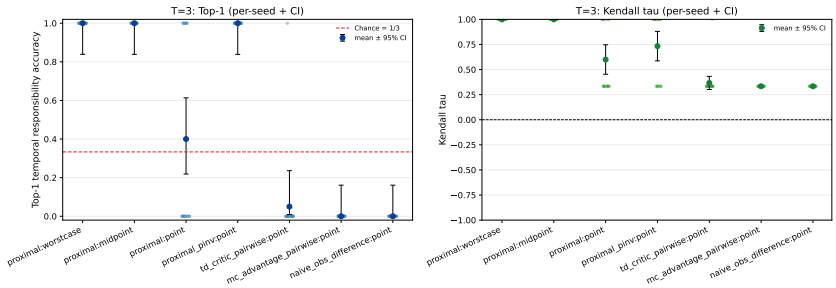

Fig. 4: **Temporal responsibility under confounding for T=3.** Left panel (Top-1 responsibility accuracy, chance = 1/3): Wilson 95% CI across seeds. Right panel (Kendall τ): mean ± 1.96·SEM across seeds.

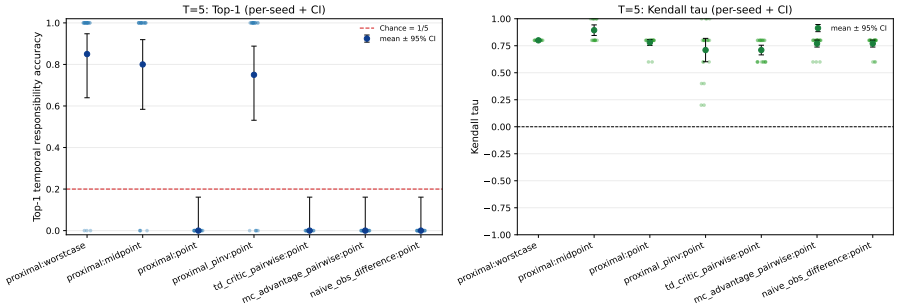

Fig. 5: **Temporal responsibility under confounding for T=5.** Left panel (Top-1 responsibility accuracy, chance = 1/5): Wilson 95% CI across seeds. Right panel (Kendall τ): mean ± 1.96·SEM across seeds. Appendix C provides timestep-selection histograms and a no-confounding sanity check.

residual and the absolute error of the point credit estimate. This establishes ρ as a highly reliable, mathematically sound threshold for autonomous agents to trigger an "abstain" or "request more data" protocol when identifiability risk is severe. Furthermore, proxy-violation ablations (Figure 7) reveal a critical insight: when SCM proxy conditional independencies are forcibly broken, interval widths can deceptively shrink even as absolute attribution error skyrockets. Therefore, narrow proximal intervals cannot act as a definitive correctness certificate if the upstream causal graph is structurally misspecified.

8.5 Multi-Agent Confounding and Algorithmic Fairness

We emphasize that our goal in these stress tests is not to evaluate the deep function approximation capabilities of modern MARL algorithms. Rather, as detailed in Appendix E, by restricting baselines like COMA and Shapley to exact tabular representations, we intentionally isolate and evaluate the fundamental

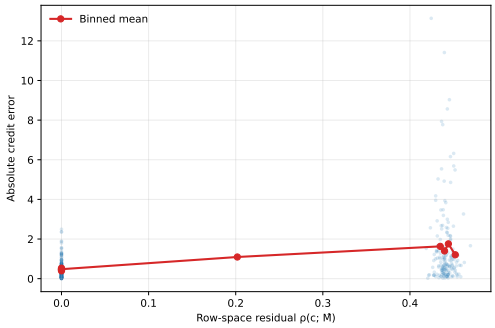

Fig. 6: **Row-space residual versus absolute credit error.** Larger residuals tend to correspond to larger credit errors, suggesting $\rho(c; \widehat{M})$ can serve as an observable warning signal.

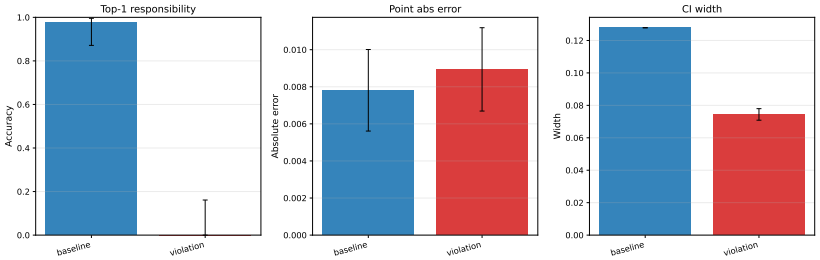

Fig. 7: **Proxy-violation ablation.** Breaking a proxy condition destroys temporal responsibility accuracy even though interval width can shrink; narrow intervals are not a correctness certificate when assumptions fail.

vulnerability of their core attribution objectives to unmeasured confounding, separated from any confounding effects of neural network optimization.

In the multi-agent stress tests (Figure 8), we established fairness-controlled comparisons. Crucially, to ensure fair evaluation, we define **matched-observables** baselines, where standard algorithms (like COMA and Shapley estimators) are provided the exact same extended observable inputs $(Z_t, W_t, \phi(H_t))$ as our proximal method. This contrasts with **baseline-standard** inputs which only receive the standard state observation defined in prior literature.

Across both the two-agent and three-agent Dec-POMDP tabular regimes, representative MARL credit baselines illustrate severe absolute error degradation under high latent confounding, regardless of whether they receive standard or matched-observable inputs. Crucially, the signed bias evaluation reveals that these correlation-based algorithms exhibit high Sign Error Rates, actively penalizing agents for cooperative actions that were actually beneficial, purely due to latent overlap with the confounder. The proximal interval credit frameworks, anchored by the moment restrictions, effectively filter the spurious noise, maintaining tightly

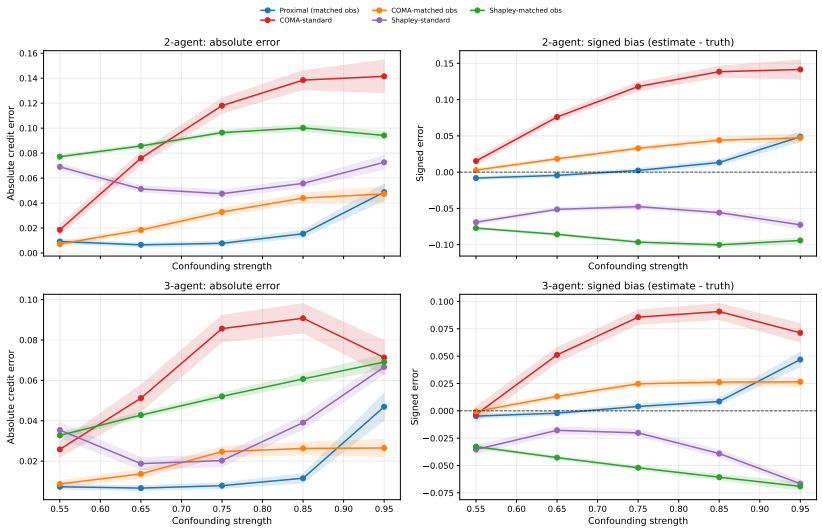

Fig. 8: **Multi-agent credit under latent confounding (two and three agents).** We report both matched-observables comparisons (fair input parity) and baseline-standard comparisons. Under stronger confounding, MARL credit baselines exhibit large absolute error and systematic bias/sign errors, while proximal interval credit is substantially more stable.

bounded absolute error and near-zero signed bias regardless of confounding severity.

9 Discussion and Limitations

Our results support a credit-first interpretation of proximal causal identification in sequential settings. When confounding is present and identification proceeds through moment restrictions, the recoverable object is not the bridge function itself, but the subset of functionals residing in the row space of the moments. Temporal credit frequently lies outside this set even when value does not.

Implications for Agent Design. When multi-agent systems act on attributions to execute debugging, adaptation, or blame allocation, structural uncertainty about credit must be explicitly propagated. A practical workflow for safe causal RL involves: (i) defining credit queries with explicit continuation policies, (ii) computing the residual-based identifiability risk $\rho(c; \widehat{M})$, (iii) reporting interval semantics or electing to abstain when the risk is unacceptably high, and (iv) executing proxy-violation falsification checks to avoid over-trusting narrow bounds.

Limitations. Our analysis is conditional on explicit discrete model classes where proximal moment equations are valid by construction, and where the bounded bridge class (Assumption 2) ensures partial identification possesses

finite semantics. We do not claim proxy assumptions are universally testable in arbitrary sequential systems, nor do we assert scalability to high-dimensional continuous environments without invoking regularized functional approximations. Our multi-agent experiments are illustrative stress tests in restricted tabular environments rather than comprehensive, deep MARL benchmarks.

10 Conclusion

Temporal credit assignment under confounding is not merely a variance reduction problem; it is fundamentally an issue of structural identifiability. In this paper, we introduced a credit-first framework, providing a formal row-space criterion to determine when temporal credit is well-posed. We rigorously proved the existence of a value-credit separation—where policy value can be identified while per-timestep attribution remains ambiguous. By proposing computable residual diagnostics, stress-testing MARL baseline vulnerabilities, and utilizing interval semantics, we provide a methodology for deploying agents that transparently represent irreducible ambiguity. Future work must focus on scaling these linear constraints to continuous Dec-POMDPs, ensuring that multi-agent systems can operate safely, interpretably, and cooperatively in real-world deployments.

References

1. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, 2nd edition (2009)
2. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, 2nd edition (2018)
3. Miao, W., Geng, Z., Tchetgen Tchetgen, E.J.: Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika, 105(4), 987–993 (2018)
4. Tchetgen Tchetgen, E.J., Ying, A., Cui, Y., Shi, X., Miao, W.: An introduction to proximal causal learning. arXiv preprint arXiv:2009.10982 (2020)
5. Cui, Y., Pu, H., Shi, X., Miao, W., Tchetgen Tchetgen, E.J.: Semiparametric proximal causal inference. Journal of the American Statistical Association, 119(546), 1348–1359 (2024)
6. Bennett, A., Kallus, N.: Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. Operations Research, 72(3), 1071–1086 (2024)
7. Foerster, J.N., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S.: Counterfactual multi-agent policy gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2018)
8. Shapley, L.S.: A value for n-person games. In: Contributions to the Theory of Games, volume II, pp. 307–317. Princeton University Press (1953)
9. Li, J., Kuang, K., Wang, B., Liu, F., Chen, L., Wu, F., Xiao, J.: Shapley counterfactual credits for multi-agent reinforcement learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), pp. 934–942 (2021)

10. Li, J., Kuang, K., Wang, B., Liu, F., Chen, L., Fan, C., Wu, F., Xiao, J.: Deconfounded value decomposition for multi-agent reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning (ICML), pp. 12843–12856. PMLR (2022)
11. Wang, J., Zhang, Y., et al.: Shapley Q-value: A local reward approach to solve global reward games. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2020)
12. Manski, C.F.: Partial Identification of Probability Distributions. Springer (2003)
13. Chakraborty, B., Murphy, S.A.: Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Springer (2014)
14. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press (2015)
15. Wang, Z., Du, Y., Zhang, Y., Fang, M., Huang, B.: MACCA: Offline multi-agent reinforcement learning with causal credit assignment. Transactions on Machine Learning Research (TMLR) (2025)
16. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research, 22, 93–115 (2004)

Appendix

A Proof sketches

Theorem 1. If $f \in \mathrm{Range}(M^\top)$ then $f = M^\top \alpha$ for some α, and for any θ with $M\theta = y$ we have $f^\top \theta = \alpha^\top y$ (unique). If $f \notin \mathrm{Range}(M^\top)$ then there exists $\delta \in \mathrm{Null}(M)$ with $f^\top \delta \neq 0$, so $f^\top(\theta + t\delta)$ varies with t across solutions.

B Implementation details: bounds, slack, SVD, and LPs

Intervals are computed by solving two linear programs ($\min / \max c^\top \theta$) over $\{\theta \in \Theta(B) : \|\widehat{M}\theta - \widehat{y}\|_\infty \leq \varepsilon_i\}$. We select the per-row slack variable via a conservative concentration bound for estimated moments (a union bound over moment rows with Hoeffding-style control for discrete probabilities/means). Specifically, we set $\varepsilon_i = B\sqrt{\log(2m/\delta)/(2\max(n_i, 1))}$, where m is the number of constraints, n_i is the number of samples for row i, and $\delta = 0.05$. We verified that our empirical structural plateau persists across varying conservative choices of ε, confirming the ambiguity is not an artificial floor created by the slack variable.

SVD for Residual Diagnostic: In finite samples, to compute a numerically meaningful row-space residual $\rho(c; \widehat{M})$, we apply a truncated SVD to the empirical moment matrix $\widehat{M} = U\Sigma V^\top$. We retain right singular vectors corresponding to singular values where $\sigma_i/\sigma_1 \geq \tau_{svd}$, and project c onto the span of these corresponding singular vectors. We verified that our empirical residual patterns are stable for standard threshold choices (e.g., $\tau_{svd} = 10^{-4}$).

C Temporal responsibility diagnostics

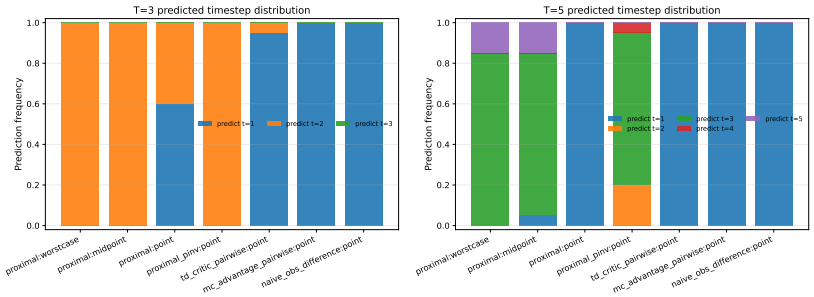

Fig. 9: **Timestep-selection histograms for $T = 3$ and $T = 5$.** Near-zero Top-1 in Figures 4 and 5 corresponds to consistent selection of an incorrect timestep (systematic misattribution), not an evaluation artifact.

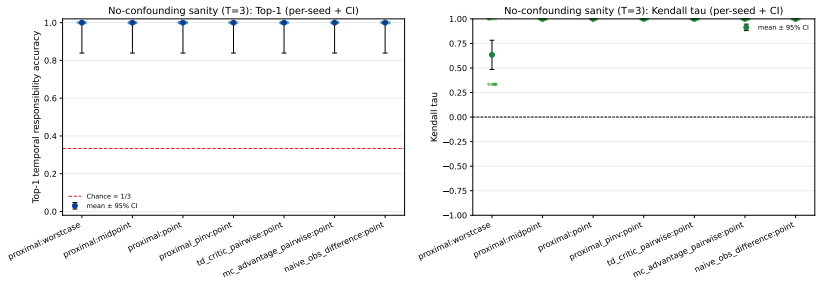

Fig. 10: **No-confounding sanity check.** When the confounder is removed, correlation-based baselines recover the correct responsible timestep, validating the evaluation pipeline.

D Overlap stress test

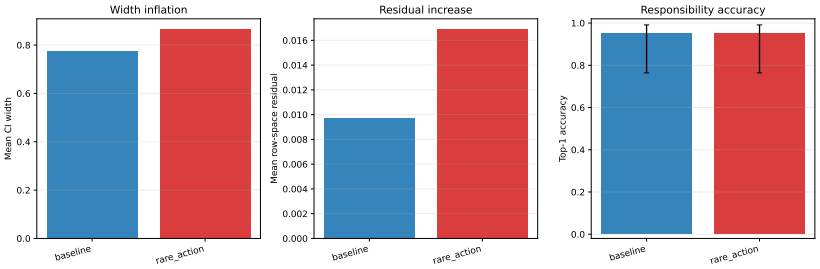

Fig. 11: **Overlap stress test.** Making an action rare inflates interval width and residual; responsibility remains high under this toy parameterization (not a general guarantee under severe overlap failure).

E Baseline Implementation Details for Tabular SCM

Deep MARL algorithms like COMA [7] and Shapley Counterfactual Credits [9] are typically deployed in high-dimensional continuous environments using deep function approximators. To evaluate them in our discrete tabular SCM ($T = 3$ and $T = 5$) without artificially crippling their representation capacity, we parameterized their centralized critics using minimal, exact-representation network architectures. Specifically, the critics were implemented using one-hot encoded state-action inputs fed into a single-layer linear network without non-linear activation functions, effectively reducing them to exact tabular updates. This ensures that any observed attribution failure stems strictly from the method's

fundamental vulnerability to unmeasured confounding, rather than a failure of deep neural network convergence in a low-data regime.