INTERPRETING EMERGENT EXTREME EVENTS IN MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model-powered multi-agent systems are powerful tools for simulating complex human-like systems. Their interactions often produce extreme events whose origins remain obscured by the black box of emergence. Interpreting these events is critical for system safety. To our knowledge, this paper proposes the first unified framework for retrospectively interpreting emergent extreme-event risk across time, agents, and behaviors in LLM-powered social simulations, answering three questions: *When does the event originate? Who drives it? And what behaviors contribute to it?* Specifically, we adapt the causal-order-constrained Shapley value to attribute reference-relative extreme-event risk to each agent action at each time step, *i.e.*, scoring the action by its marginal contribution in replay. We then aggregate the attribution scores over time, agents, and behaviors to quantify the risk contribution of each dimension. Finally, we design metrics from these contribution scores to characterize extreme events. Experiments across diverse multi-agent system scenarios (economic, financial, and social) demonstrate how our framework characterizes extreme events through their temporal, agent, and behavioral contributions.

1 INTRODUCTION

Large language models (LLMs) can capture human behavioral patterns in language, decision-making, and social interactions (Grossmann et al., 2023; Xi et al., 2025). These capabilities enable highly human-like multi-agent systems (MAS) that consist of multiple LLM-powered agents. Recent studies use LLM-powered MAS to simulate complex systems (Park et al., 2023; Li et al., 2024). Here, LLM-powered agents interact autonomously with each other and the environment, producing emergent complex phenomena.

> *"A small number of Black Swans explain almost everything in our world."*
>
> — Nassim Nicholas Taleb, *The Black Swan*

Among the phenomena emerging from MAS, we focus on ***Black Swans*** (Taleb, 2007). As illustrated in Figure 1(a), Black Swans refer to extreme events with three key properties: (1) they are *outliers* falling outside regular expectations, typically residing in the tail of a fat-tailed distribution (Clauset et al., 2009; Taleb, 2020); (2) they carry an *extreme impact*, *e.g.*, causing systemic collapse or drastic state transitions; and (3) they feature *retrospective predictability*, meaning that explanations are constructed after the event. Existing literature has documented related phenomena in MAS, including soaring inflation in economic systems (Li et al., 2024), market crashes in financial markets (Yang et al., 2025), and group polarization in social networks (Wang et al., 2025).

However, interpreting emergent extreme events in MAS remains a critical challenge. Complex agent-environment interactions create a black box of emergence (Miller, 2017; Baker et al., 2020). The factors driving the disaster across dimensions remain hidden, leaving fundamental questions underexplored: ***When does the extreme event originate? Who drives it? And what behaviors contribute to it?*** Interpreting the extreme events provides both theoretical analysis of the underlying mechanisms and a basis for identifying high-risk agents and behaviors. However, a unified interpretation across time, agents, and behaviors remains underexplored in these simulations.

To our knowledge, we propose the first unified framework to interpret emergent extreme-event risk across time, agents, and behaviors in LLM-powered social simulations. As shown in Figure 1(b), our

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.

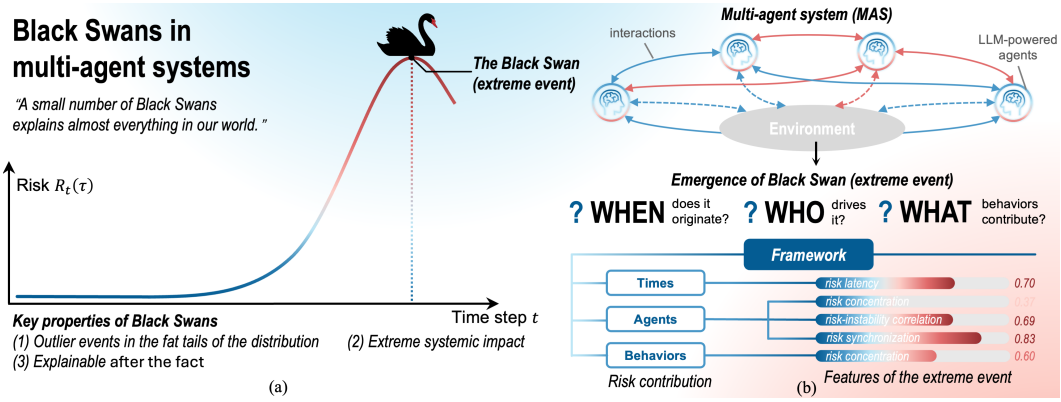


Figure 1: Interpreting Black Swans in multi-agent systems. (a) Rare, high-impact events admit retrospective explanations. (b) Our framework attributes emergent risk across time, agents, and behaviors, answering *when*, *who*, and *what*. Displayed metric values are illustrative.

framework answers *when*, *who*, and *what*. We analyze recorded extreme events to identify the key temporal phases, agents, and behaviors contributing to reference-relative risk. Specifically, as Figure 2 shows, we assign an attribution score to each agent's action to quantify its event contribution using the causal-order-constrained Shapley value (Shapley, 1953; Frye et al., 2020). We then aggregate action-level scores across time steps, agents, and behavior types to quantify the contributions of each dimension. These aggregates yield metrics characterizing extreme events.

We applied our framework to three representative MAS scenarios: economic systems, financial markets, and social networks. For instance, the selected financial cases attribute risk close to the event, identify a short-term investor with a large positive contribution, and separate trading contributions across indices. More broadly, our analysis across diverse scenarios yields five insights (I1–I5) about extreme events in LLM-powered MAS, spanning the *when*, *who*, and *what* dimensions. The following insights summarize the attribution patterns observed in our simulations, with metrics defined in Section 3 and supporting results discussed in Section 4.3.

Insight 1. Extreme-event risk can be attributed to early contributions or to actions close to the event.
Insight 2. Risk contributions tend to concentrate among a subset of agents.
Insight 3. Agents with larger absolute risk contributions tend to exhibit greater temporal variability.
Insight 4. Agents' risk contributions are partially aligned in direction.
Insight 5. Risk contributions tend to concentrate in a few behavior types.

2 Related Work

LLM-powered MAS for simulations: Since LLMs can capture human reasoning and complex social interaction behaviors (Wei et al., 2022; Park et al., 2023; Zhao et al., 2024), multi-agent systems (MAS) composed of these models are widely used for human-like simulations (Park et al., 2023; Li et al., 2023; Yang et al., 2024; Li et al., 2024; Yang et al., 2025; Wang et al., 2025). We focus on *social-simulation MAS*, which model emergent economic, financial, and social phenomena without a predefined correct outcome, as distinct from *task-completion MAS* with defined goals (Zhang et al., 2025a; Cemri et al., 2025). These simulations also reproduce phenomena such as rising inflation (Li et al., 2024), market crashes (Yang et al., 2025), and opinion polarization (Wang et al., 2025). However, reproducing these events leaves open how they arise from individual actions; our work addresses this interpretive gap.

Attribution and risk analysis in MAS: Attribution methods interpret black-box models by quantifying input contributions to their outputs (Lundberg & Lee, 2017; Sundararajan et al., 2017). Recently, these techniques have been adapted to MAS. For instance, Ge et al. (2025) estimate action suspiciousness through repeated execution and spectrum analysis; Zhang et al. (2025a) train a failure-localization model using counterfactual replay and fault-injection data; and Zhang et al. (2025b); Cemri et al. (2025) use LLM prompting for failure attribution or taxonomy-based diagno-

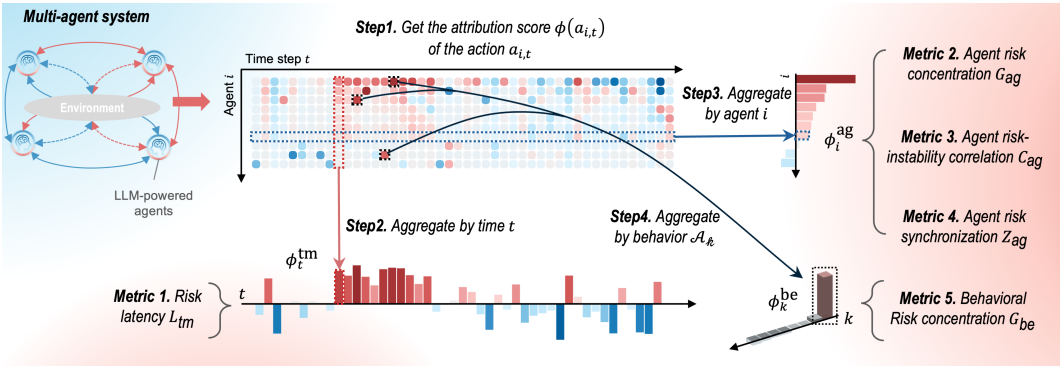


Figure 2: Causal-order-constrained Shapley attribution (Step 1) and its aggregation into temporal, agent, and behavioral metrics (Steps 2–4). Red/blue indicate positive/negative reference-relative contributions.

sis. These methods primarily address task failures. Risk-sensitive MARL instead optimizes policies under risk constraints (Shen et al., 2023; Slumbers et al., 2023). Related work also applies Shapley attribution to road-safety responsibility (Niu et al., 2024), decomposes counterfactual effects through subsequent agent actions (Triantafyllou et al., 2025), and explains collective behavior using structural causal models (Weinberg, 2026). Our question is: how recorded actions contribute to emergent extreme-event risk in social simulations across *when*, *who*, and *what*.

Quantitative analysis of extreme events: Previous work quantifies extreme events from different perspectives. For example, Gabaix (2009) studies power-law scaling in economic and financial variables, while Scheffer et al. (2009); Dakos et al. (2012) examine early-warning indicators such as autocorrelation and variance. These system-level summaries do not assign contributions to individual actions. Our attribution-based metrics interpret extreme events through *when*, *who*, and *what*.

3 Method

Our framework analyzes extreme events in MAS through three questions: *When does the extreme event originate? Who drives it? And what behaviors contribute to it?* We formulate extreme event attribution with causal-order-constrained Shapley values, aggregate the scores along time, agents, and behaviors, and design interpretable metrics from these aggregates. Each trajectory contains N agents over a common horizon T, with one recorded action instance per agent and step. An instance may bundle several decisions or represent a no-operation action.

3.1 Preliminaries: Extreme Event Attribution in MAS

We model the LLM-powered MAS as a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P \rangle$ with N agents in $\mathcal{N} = \{1, \ldots, N\}$, global state space $\mathcal{S}$, action space $\mathcal{A}$, and transition function P. At step t, each agent i executes $a_{i,t} \in \mathcal{A}$ from state $s_t \in \mathcal{S}$, and the joint actions induce $s_{t+1} \sim P(\cdot \mid s_t, \{a_{i,t}\}_{i=1}^N)$, producing a trajectory $\tau = (s_1, \{a_{i,1}\}_{i=1}^N, s_2, \ldots, \{a_{i,T}\}_{i=1}^N, s_{T+1})$, where T counts action steps up to the selected event endpoint. The risk indexed by T is evaluated after these actions, at state s_{T+1}.

Problem Formulation. To quantify an extreme event in trajectory τ, we introduce real-valued raw-risk metrics $\{\hat{R}_t(\tau) \mid t = 1, \ldots, T\}$. Let $\tau^{\emptyset}$ denote the all-safe baseline trajectory and define the attributable excess risk as $R_t(\tau) := \hat{R}_t(\tau) - \hat{R}_t(\tau^{\emptyset})$. For each selected extreme-event trajectory, we fix T at the historical peak of its system-specific raw-risk indicator, using the recorded selection protocol (Appendix B). Event selection uses factual measurements; the reference-relative contrast need not be positive. Coalition replays retain this endpoint instead of selecting a new peak. We assign scores $\phi(a_{i,t})$ to decompose its excess risk:

$$R_T(\tau) = \sum_{t=1}^{T} \sum_{i=1}^{N} \phi(a_{i,t}). \tag{1}$$

Table 1: Cosine similarity to exact constrained attributions in six-player synthetic games. Mean ± sample standard deviation over 10 independent runs; budgets are nested prefixes.

Game	$\lvert\Pi_\prec\rvert$	$M = 10$	$M = 100$	$M = 300$	$M = 1{,}000$
Chain	1	1.00000 ± 0.00000	1.00000 ± 0.00000	1.00000 ± 0.00000	1.00000 ± 0.00000
Branches	16	0.99902 ± 0.00121	0.99990 ± 0.00009	0.99995 ± 0.00004	0.99999 ± 0.00001
Empty order	720	0.99114 ± 0.01185	0.99899 ± 0.00107	0.99972 ± 0.00023	0.99990 ± 0.00008
Time blocks	36	0.99336 ± 0.00517	0.99931 ± 0.00052	0.99982 ± 0.00017	0.99993 ± 0.00005

Reference choice is part of this problem definition. Reference-based attribution explains a contrast between an observed outcome and a reference outcome (Sundararajan et al., 2017; Sundararajan & Najmi, 2020). In our setting, the reference uses fixed labor and consumption in EconAgent, holding positions in TwinMarket, and no posting or interaction in SocialNetwork. We retain the term "safe reference" for this prescribed comparison, without assuming that it minimizes risk. Different references answer different questions; Appendix D.3 evaluates their effect on action rankings.

Attributing the extreme event using causal-order-constrained Shapley values. The Shapley value was originally introduced in game theory to allocate a game's total reward among its players (Shapley, 1953). To incorporate causal precedence among actions, we adopt the asymmetric Shapley framework (Frye et al., 2020), restricting the orders in which actions enter a coalition. For a fixed precedence relation, the allocation satisfies *efficiency*, *nullity*, and *linearity*; symmetry holds under relabelings that preserve both the game and its constraints. Detailed proofs are provided in Appendix A.1. In a sequential simulation, retaining an action while replacing a required predecessor violates the supplied dependency. We therefore evaluate marginal contributions only in predecessor-closed coalitions: including an action requires including its specified predecessors.

We use causal-order-constrained Shapley values to calculate the attribution score $\phi(a_{i,t})$. Given trajectory τ, the players are all recorded action instances $\Omega = \{a_{i,t} : i \in \mathcal{N},\, 1 \le t \le T\}$, where $n = |\Omega| = NT$. Let τ^S denote a counterfactual replay that preserves actions in $S \subseteq \Omega$ and replaces the others with reference actions $\tilde{a}_{i,t}$, while states evolve under the simulator dynamics. Its characteristic function is the excess risk $v(S) = \hat{R}_T(\tau^S) - \hat{R}_T(\tau^\emptyset)$, so $v(\emptyset) = 0$ and verified full replay gives $v(\Omega) = R_T(\tau)$. The endpoint, reference, and replay randomness remain fixed. Let $\prec$ be an acyclic precedence relation derived from simulator mechanisms and recorded information flow, and let $\Pi_\prec(\Omega)$ contain the permutations respecting every constraint $a \prec b$. This partial relation is fixed before attribution. For $\pi \in \Pi_\prec(\Omega)$, write $\mathrm{Pre}_{i,t}(\pi)$ for the action instances preceding $a_{i,t}$ in π. We define:

$$\phi(a_{i,t}) = \frac{1}{|\Pi_\prec(\Omega)|} \sum_{\pi \in \Pi_\prec(\Omega)} \Big[v\big(\mathrm{Pre}_{i,t}(\pi) \cup \{a_{i,t}\}\big) - v\big(\mathrm{Pre}_{i,t}(\pi)\big) \Big]. \tag{2}$$

Every admissible prefix contains the specified predecessors of its actions. These constraints govern coalition inclusion, while replay always follows the simulator's chronology. Without constraints, Eq. 2 reduces to Ordinary Shapley, used as a baseline with the same v. The scores decompose reference-relative terminal risk under the supplied causal order and replay protocol (Appendix A.3).

Estimating Shapley values using Monte Carlo sampling. Exact computation enumerates admissible orders. We estimate their average with M sampled orders, requiring at most $1 + M|\Omega|$ replay evaluations. Complete chronological blocks allow independent within-block shuffling; general partial orders use four adjacent-swap Markov chains targeting the uniform admissible-order distribution. The latter draws are approximate and correlated. Appendix A.2 provides pseudocode and the sampling overhead.

Verifying the approximation accuracy of Monte Carlo sampling. Let $\boldsymbol{\phi}^*$ denote the exact attribution vector of an enumerable game and $\widehat{\boldsymbol{\phi}}$ its Monte Carlo estimate. We use cosine similarity $\cos(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}^*)$ to measure approximation accuracy. We test four six-player games, reporting means and sample standard deviations over 10 independently seeded runs (Table 1). At $M = 1{,}000$, mean similarity is at least 0.99990 in the three nontrivial cases, with mean relative ℓ_2 errors of 0.50%, 1.33%, and 1.12%. The single-order chain is exact by construction. These results support accurate approx-

Table 2: Causal-order-constrained Shapley metrics: mean ± sample SD. Sample counts are given in Appendix B. Model names denote archive groups (Section 4.1).

Model	When L_{tm}	G_{ag}	Who C_{ag}	Z_{ag}	What G_{be}
EconAgent					
GPT	0.531 ± 0.333	0.467 ± 0.117	0.569 ± 0.335	0.631 ± 0.050	0.711 ± 0.046
Llama	0.725 ± 0.204	0.564 ± 0.114	0.646 ± 0.340	0.816 ± 0.042	0.661 ± 0.130
Claude	0.665 ± 0.205	0.371 ± 0.080	0.623 ± 0.415	0.856 ± 0.042	0.566 ± 0.060
Qwen	0.741 ± 0.218	0.480 ± 0.078	0.634 ± 0.335	0.779 ± 0.023	0.551 ± 0.098
DeepSeek	0.831 ± 0.141	0.429 ± 0.089	0.521 ± 0.344	0.811 ± 0.044	0.639 ± 0.053
TwinMarket					
GPT	0.075 ± 0.061	0.504 ± 0.069	0.767 ± 0.194	0.371 ± 0.060	0.495 ± 0.048
Llama	0.094 ± 0.059	0.603 ± 0.117	0.919 ± 0.104	0.414 ± 0.049	0.605 ± 0.017
Claude	0.089 ± 0.046	0.650 ± 0.085	0.979 ± 0.014	0.427 ± 0.139	0.570 ± 0.101
Qwen	0.120 ± 0.121	0.545 ± 0.078	0.789 ± 0.167	0.314 ± 0.099	0.498 ± 0.087
DeepSeek	0.055 ± 0.019	0.522 ± 0.070	0.876 ± 0.067	0.381 ± 0.047	0.611 ± 0.068
SocialNetwork					
GPT	0.067 ± 0.070	0.360 ± 0.049	0.438 ± 0.177	0.356 ± 0.047	0.574 ± 0.043
Llama	0.254 ± 0.194	0.282 ± 0.043	-0.072 ± 0.344	0.264 ± 0.037	0.680 ± 0.027
Claude	0.227 ± 0.224	0.354 ± 0.037	0.218 ± 0.391	0.314 ± 0.059	0.612 ± 0.074
Qwen	0.420 ± 0.241	0.469 ± 0.100	0.570 ± 0.408	0.404 ± 0.076	0.634 ± 0.049
DeepSeek	0.171 ± 0.184	0.342 ± 0.087	0.529 ± 0.409	0.322 ± 0.064	0.633 ± 0.056

imation in these enumerable games; Appendix D.1 specifies their construction. On one fixed full trajectory per system, 20 independent $M = 1{,}000$ estimates give MRSD 4.61%, 13.01%, and 6.08% for EconAgent, TwinMarket, and SocialNetwork. The economic and social cases meet our ranking diagnostics at this budget. Averaging eight independently seeded financial runs meets these criteria (minimum Spearman 0.961 across five independent averages), motivating the sampling budgets in Section 4.1. Appendix D.2 gives the complete precision checks.

3.2 When does the Extreme Event Originate?

Evaluating the risk contribution ϕ_t^{tm} of each time step t to the extreme event. Our goal is to identify the key time step by which most of the terminal risk has been attributed. As shown in Figure 2(a), we aggregate the attribution scores of all actions at step t:

$$\phi_t^{\mathrm{tm}} = \sum_{i=1}^{N} \phi(a_{i,t}). \tag{3}$$

As defined in Section 3.1, the attributed excess risk satisfies $R_T(\tau) = \sum_{t=1}^{T} \phi_t^{\mathrm{tm}}$. For $R_T(\tau) > 0$, we define T^* as the first step at which cumulative attribution exceeds a proportion q of the terminal risk: $T^* = \min\{t' \in \{1, \ldots, T\} : \sum_{t=1}^{t'} \phi_t^{\mathrm{tm}} > qR_T(\tau)\}$. We use $q = 0.900$. This aggregation threshold affects only L_{tm}, not the action attributions or the four agent/behavior metrics. Because contributions can be signed, cumulative attribution need not be monotone: T^* records the first crossing and need not remain above the threshold thereafter.

Definition 1. ***(Metric: relative risk latency L_{tm})*** *To evaluate the earliness of attributed risk relative to the event duration, we define*

$$L_{tm} = (T - T^*)/T \in [0, 1 - 1/T], \tag{4}$$

where T^* is the first crossing of $qR_T(\tau)$ by cumulative attribution. This metric is defined for positive terminal excess risk and $0 < q < 1$.

As Figure 3 shows, a large L_{tm} indicates that substantial risk contribution is attributed to early actions, whereas a small value indicates contribution close to the event. This is a retrospective attribution measure; Appendix A.3 discusses its dependence on the precedence constraints.

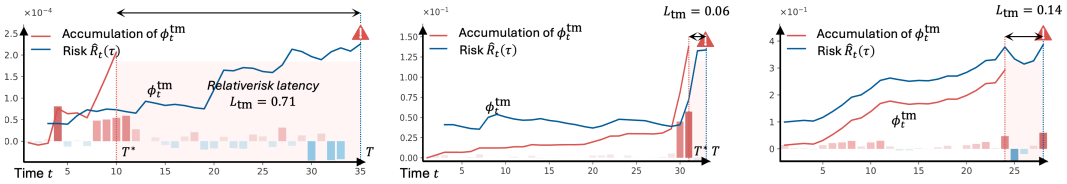


Figure 3: Temporal attributions and observed raw risk $\hat{R}_t(\tau)$: GPT seeds 52, 44, and 44 in EconAgent, TwinMarket, and SocialNetwork, respectively. Red cumulative curves are shown through T^*, the first strict 90.00% crossing of terminal excess risk.

3.3 Who Drives the Extreme Event?

Evaluating the risk contribution ϕ_i^{ag} of each agent i to the extreme event. We aggregate the agent's action scores over time and measure their temporal variability:

$$\phi_i^{\text{ag}} = \sum_{t=1}^{T} \phi(a_{i,t}), \quad \sigma_i^{\text{ag}} = \text{std}_{t=1}^{T}\left(\phi(a_{i,t})\right). \tag{5}$$

Here std measures attribution variability using the population denominator T.

Definition 2. ***(Metric: agent risk concentration G_{ag})*** *We use the Gini coefficient (Hurley & Rickard, 2009) of absolute net agent contributions:*

$$G_{ag} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left||\phi_i^{ag}| - |\phi_j^{ag}|\right|}{2N\sum_{i=1}^{N}|\phi_i^{ag}|} \in [0, 1 - 1/N]. \tag{6}$$

It is undefined when all net agent contributions vanish.

G_{ag} measures the concentration of absolute net risk contributions. A higher value indicates concentration among fewer agents, while a lower value indicates a more even distribution. Because magnitudes include protective contributions, signed scores identify risk-increasing agents.

Definition 3. ***(Metric: risk-instability correlation C_{ag})*** *We summarize the linear association of absolute net contribution and temporal attribution dispersion across agents using Pearson correlation (Rodgers & Nicewander, 1988):*

$$C_{ag} = \frac{\text{Cov}(\boldsymbol{\phi}, \boldsymbol{\sigma})}{\text{std}(\boldsymbol{\phi}) \cdot \text{std}(\boldsymbol{\sigma})} \in [-1, 1], \tag{7}$$

where $\boldsymbol{\phi} = [|\phi_1^{ag}|, \ldots, |\phi_N^{ag}|]$ and $\boldsymbol{\sigma} = [\sigma_1^{ag}, \ldots, \sigma_N^{ag}]$. Covariance and standard deviations use the same population convention over agents. Correlation is undefined if either vector is constant.

A large positive C_{ag} indicates that agents with larger absolute net contributions also exhibit greater temporal variability in their action attributions, as shown in Figure 4(a). A negative value indicates the reverse.

Definition 4. ***(Metric: agent risk synchronization Z_{ag})*** *As a scalar analogue of normalized alignment in the Vicsek order parameter (Vicsek et al., 1995), we average per-step sign alignment. A zero denominator gives a zero summand, retaining that step in the average over all T steps:*

$$Z_{ag} = \frac{1}{T}\sum_{t=1}^{T}\frac{\left|\sum_{i=1}^{N}\phi(a_{i,t})\right|}{\sum_{i=1}^{N}|\phi(a_{i,t})|} \in [0, 1]. \tag{8}$$

A high Z_{ag} indicates that agents' contributions tend to have the same sign within a step, while a low value reflects cancellation between positive and negative contributions or steps with zero attribution.

3.4 What Behaviors Contribute to the Extreme Event?

Evaluating the risk contribution ϕ_k^{be} of each action pattern $\mathcal{A}_k$ to the extreme event. Let $w_{i,t,k} \geq 0$ denote action membership in behavior category $\mathcal{A}_k$, using the scenario-specific mapping in Appendix B. Single-category actions have unit weight; composite actions may split their

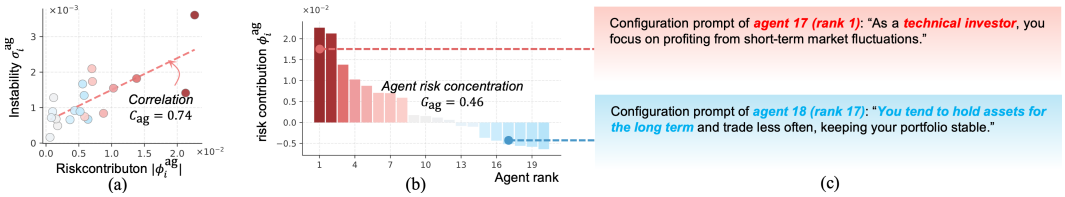


Figure 4: Agent-level attribution for GPT seed 45 in TwinMarket. Bars are sorted by signed contribution; agent labels use fixed indices. Initial configuration excerpts are translated from the logged Chinese prompts.

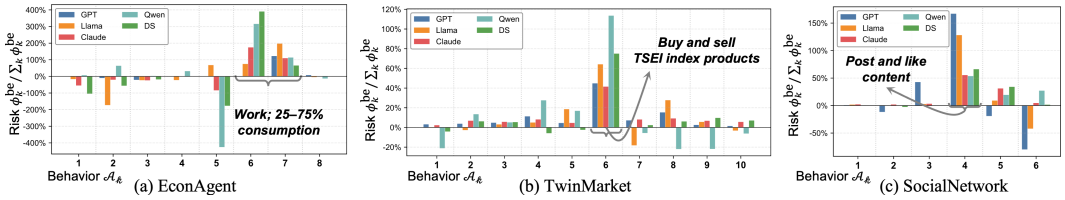


Figure 5: Mean normalized behavior contributions across scenarios and models. Signed shares $\phi_k^{\text{be}}/R_T(\tau)$ are averaged over trajectories with positive terminal excess risk; they may be negative or exceed 100.00%.

contribution across categories:

$$\phi_k^{\text{be}} = \sum_{t=1}^{T}\sum_{i=1}^{N} w_{i,t,k}\,\phi(a_{i,t}). \tag{9}$$

The weights sum to one for each represented action. No-order/no-interaction actions with no category assignment are individually checked to have zero attribution. Thus $\sum_k \phi_k^{\text{be}} = R_T(\tau)$. These scores measure total category contributions; dividing by the effective occurrence count $c_k = \sum_{t,i} w_{i,t,k}$ gives a per-occurrence score for $c_k > 0$ (Appendix D.7).

Definition 5. ***(Metric: behavior risk concentration G_{be})*** *To measure risk concentration across behavior types, we use the Gini coefficient (Hurley & Rickard, 2009) of the absolute aggregated contributions across the K categories:*

$$G_{be} = \frac{\sum_{s=1}^{K}\sum_{k=1}^{K}\big||\phi_s^{be}| - |\phi_k^{be}|\big|}{2K\sum_{k=1}^{K}|\phi_k^{be}|} \in [0,1]. \tag{10}$$

It is undefined when all aggregated contributions vanish.

A high G_{be} indicates that contribution magnitudes concentrate in a few behavior types, whereas a low value indicates a more even distribution. Signed scores distinguish risk-increasing from protective contributions.

4 Experiments

4.1 Set Up

We evaluate five model groups (GPT, Llama, Claude, Qwen, and DeepSeek) in three MAS scenarios: **EconAgent** (Li et al., 2024) (inflation risk), **TwinMarket** (Yang et al., 2025) (market risk), and **SocialNetwork** (polarization). We select trajectories exhibiting extreme events under the scenario-specific criteria; Appendix B provides the selection details and sample counts. We report means and sample standard deviations across trajectories. Action spaces, risk metrics, and event selection are described in Appendix B. Full-action replay reproduces the recorded risk endpoints. Both Shapley methods use $M = 1{,}000$ in EconAgent and SocialNetwork. In TwinMarket, each method averages eight independently seeded $1{,}000$-permutation estimates; metrics are recomputed from the mean action vector. Appendix D.2 reports the fixed-case precision checks. Figure 5 uses trajectories with positive terminal excess risk. Figure 5 averages their signed normalized behavior contributions without outlier filtering. Model names denote archive groups; Appendix B gives generation metadata.

Table 3: Intervention-grounded recovery: means over 144 dependent games and 36 independent controls. Err.: normalized recovery error (%); Top-1: accuracy (%); r_s: Spearman correlation. FT/FA use DeepSeek-V3.2. Graph controls alter only ASV's supplied graph. Standard deviations and protocol are in Appendix D.10.

	Dependent games			Independent controls		
Method	Err.↓	Top-1↑	r_s ↑	Err.↓	Top-1↑	r_s ↑
SV (Shapley, 1953)	30.09	36.81	0.629	0.00	100.00	1.000
LOO (Covert et al., 2021)	26.41	34.72	0.627	0.00	100.00	1.000
FT (Cemri et al., 2025)	60.16	20.01	0.090	37.58	43.84	0.391
FA (Zhang et al., 2025b)	75.52	11.59	-0.087	48.41	29.14	0.220
ASV, half-edge retention	19.75	48.61	0.736	0.00	100.00	1.000
ASV, relabelled graph	34.03	38.89	0.569	0.00	100.00	1.000
ASV (ours) (Frye et al., 2020)	0.00	100.00	1.000	0.00	100.00	1.000

4.2 Effectiveness Evaluation

Compared methods. SV (Shapley, 1953) averages marginal contributions over all action permutations. **LOO** (Covert et al., 2021) assigns the full-minus-single-action contrast $v(\Omega) - v(\Omega \setminus \{a\})$. **FT** (Cemri et al., 2025) uses failure-taxonomy prompting; **FA** (Zhang et al., 2025b) uses sequence-based attribution prompting. We adapt both to score every action for the specified risk target, using matched inputs and model settings. **ASV (ours)** (Frye et al., 2020) adapts asymmetric Shapley values to the replay game, averaging only over orders consistent with the supplied causal prerequisites. Appendix C describes the baselines.

Intervention-grounded evaluation. We construct 180 risk-accumulation games with 6, 8, or 10 actions, spanning chains, forks, joins, random DAGs, and independent controls. An event occurs only when its action and all prerequisite events occur; each realized event adds a fixed risk increment. For each action, paired simulator interventions complete its ancestors, keep other requests at reference, and measure the risk added by executing that action. This supplies an independent stage-local contribution target. Table 3 reports normalized recovery error, Top-1 accuracy, and rank correlation. SV, LOO, and ASV are exact; FT/FA receive the complete coalition lookup with coefficients and target labels withheld. Appendix D.10 gives the protocol and recovery property; Appendix D.9 reports additional structural diagnostics.

ASV recovers the intervention-derived contributions to numerical precision in all 144 dependent games. Removing or perturbing graph edges increases its error, demonstrating the role of correct causal structure in this mechanism class. Without dependencies, the three exact methods agree. FT/FA show larger recovery errors under the matched prompting protocol.

Computational profile. Attribution uses $O(MNT)$ replay evaluations, plus graph and sampling overhead. Appendix D.6 separates attribution, generation, and LLM-baseline costs.

4.3 When, Who, and What: Main Results

When: early and near-event contributions (I1). Table 2 and Figure 3 connect the timing summary to individual action contributions. A large L_{tm} places the cumulative-attribution crossing early in the recorded event, while a small value places it close to the endpoint. The illustrated trajectories show both patterns. Their per-step bars retain positive and negative contributions after the first crossing, revealing how later actions amplify or offset earlier contributions.

Who: concentration, variability, and alignment (I2–I4). Table 2 shows uneven agent contributions, with mean G_{ag} between 0.282 and 0.650. Mean C_{ag} is positive in 14 of 15 settings, supporting the tendency for agents with larger absolute contributions to exhibit greater temporal variability. Mean Z_{ag} ranges from 0.264 to 0.856, indicating partial directional alignment. Figure 4 illustrates the signed contributions: short-term investor agent 17 has the largest positive net contribution, whereas long-term holder agent 18 contributes negatively in this trajectory.

Table 4: Additional EconAgent subset metrics (top; mean $\pm$ sample SD) and fixed-trajectory q sensitivity (bottom; one trajectory).

Subset cutoff (size)	L_{tm}	G_{ag}	C_{ag}	Z_{ag}	G_{be}
$0.800\rho_0$ ($n = 25$)	0.683 ± 0.177	0.472 ± 0.139	0.575 ± 0.383	0.800 ± 0.074	0.630 ± 0.101
ρ_0 ($n = 19$)	0.674 ± 0.190	0.476 ± 0.151	0.527 ± 0.412	0.812 ± 0.070	0.651 ± 0.095
$1.200\rho_0$ ($n = 15$)	0.630 ± 0.185	0.487 ± 0.156	0.553 ± 0.435	0.818 ± 0.068	0.660 ± 0.100
Proportion q	L_{tm}	G_{ag}	C_{ag}	Z_{ag}	G_{be}
0.800	0.607	0.551	0.856	0.694	0.741
0.900 (default)	0.500	0.551	0.856	0.694	0.741
0.950	0.464	0.551	0.856	0.694	0.741

What: concentrated behavior contributions (I5). Mean G_{be} ranges from 0.495 to 0.711 in Table 2, reflecting uneven contribution magnitudes across behavior categories. Figure 5 identifies the prominent categories and separates positive from protective contributions. Together with the temporal and agent views, these summaries connect an extreme event to its contributing actions.

4.4 Parameter Sensitivity Analysis

We examine two parameters with distinct roles: the excess-risk cutoff ρ changes the set of trajectories being summarized, whereas the cumulative-attribution threshold q changes the reported risk latency. Table 4 separates these analyses. **Sensitivity to the event subset.** In EconAgent, we retain trajectories with terminal excess risk $v(\Omega) \geq \rho$, using $\rho \in \{0.800\rho_0, \rho_0, 1.200\rho_0\}$ and the frozen median positive excess risk $\rho_0 \approx 1.655 \times 10^{-4}$. These subsets contain 25, 19, and 15 trajectories. Across these cutoffs, mean agent concentration G_{ag} ranges from 0.472 to 0.487, and mean behavior concentration G_{be} from 0.630 to 0.660. Mean latency L_{tm} decreases from 0.683 to 0.630. Concentration therefore changes little across these subsets, while the timing summary shifts with event selection. Each trajectory's attribution is fixed across cutoffs.

Sensitivity to the latency threshold. On one fixed EconAgent trajectory, increasing q from 0.800 to 0.950 moves the first crossing from step 11 to step 15 and reduces L_{tm} from 0.607 to 0.464; the default $q = 0.900$ gives $L_{\mathrm{tm}} = 0.500$. The four agent and behavior metrics remain unchanged because q enters only the latency calculation. Thus q controls the attributed risk required for the crossing, without altering the action-level allocation. This check and the subset analysis use different trajectories; Appendix D.4 gives their construction. Appendices D.3 and D.8 report reference-action and precedence sensitivity.

5 Conclusion

Our framework interprets emergent extreme events in LLM-powered multi-agent systems through three questions: *when does the event originate, who drives it, and what behaviors contribute?* We adapt the asymmetric Shapley framework to allocate reference-relative terminal risk among recorded agent actions using dependencies specified by the simulator. Aggregating these action-level contributions over time, agents, and behaviors yields a unified explanation of the event, with five metrics characterizing its temporal development, agent concentration, variability, direction alignment, and behavioral concentration. Experiments in economic, financial, and social simulations characterize these contributions in recorded extreme events. The framework links system-level extreme events to their underlying actions.

Limitations. Computational cost remains the main limitation: exact attribution averages over all admissible orders, requiring repeated simulation. We use Monte Carlo estimation, with exact within-block shuffling for chronological constraints and approximate Markov-chain sampling for general partial orders. Small-game comparisons and repeated-run checks guide our sampling budgets.

AI USE STATEMENT

Generative AI tools assisted with writing and language polishing, literature retrieval and related-work discovery, research ideation and execution, and drafting sections of the manuscript. Research assistance included methodological discussions and the implementation, debugging, and execution of attribution experiments. The authors take responsibility for the final manuscript, code, and results. LLM-based simulation agents and attribution baselines are described in Appendices B and C.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A.1 specify the core attribution algorithms and their assumptions. Appendices A.2–B describe the sampling procedures, simulation settings, and replay protocols, while Appendix D provides the experimental details. We will provide an anonymized implementation of the core algorithms as supplementary material to support reproducibility.

REFERENCES

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula, 2020. URL https://arxiv.org/abs/1909.07528.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL http://jmlr.org/papers/v22/20-1316.html.

Vasilis Dakos, Stephen R Carpenter, William A Brock, Aaron M Ellison, Vishwesha Guttal, Anthony R Ives, Sonia Kéfi, Valerie Livina, David A Seekell, Egbert H van Nes, et al. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS one*, 7(7):e41010, 2012.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html.

Xavier Gabaix. Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294, 2009.

Yu Ge, Linna Xie, Zhong Li, Yu Pei, and Tian Zhang. Who is introducing the failure? automatically attributing failures of multi-agent systems via spectrum analysis. *arXiv preprint arXiv:2509.13782*, 2025.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. Ai and the transformation of social science research. *Science*, 380(6650): 1108–1109, 2023.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15523–15536, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *ArXiv*, abs/1706.07269, 2017. URL https://arxiv.org/abs/1706.07269.

Luyao Niu, Hongchao Zhang, Dinuka Sahabandu, Bhaskar Ramasubramanian, Andrew Clark, and Radha Poovendran. Who is responsible? explaining safety violations in multi-agent cyber-physical systems. *arXiv preprint arXiv:2410.20288*, 2024. URL https://arxiv.org/abs/2410.20288.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Siqi Shen, Chennan Ma, Chao Li, Weiquan Liu, Yongquan Fu, Songzhu Mei, Xinwang Liu, and Cheng Wang. RiskQ: Risk-sensitive multi-agent reinforcement learning value factorization. In *Advances in Neural Information Processing Systems*, 2023.

Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32059–32087. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/slumbers23a.html.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007. ISBN 9781400063512.

Nassim Nicholas Taleb. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*, 2020.

Stelios Triantafyllou, Aleksa Sukovic, Yasaman Zolfimoselo, and Goran Radanovic. Counterfactual effect decomposition in multi-agent sequential decision making. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60072–60098. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/triantafyllou25a.html.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. In *Proceedings of the 31st international conference on computational linguistics*, pp. 3913–3923, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Abraham Itzhak Weinberg. Macie a multi-agent causal intelligence explainer for collective behavior understanding. *Discover Artificial Intelligence*, 6:566, 2026. doi: 10.1007/s44163-026-01543-2.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. *arXiv preprint arXiv:2502.01506*, 2025.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. Agentracer: Who is inducing failure in the llm agentic systems? *arXiv preprint arXiv:2509.03312*, 2025a.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, et al. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*, 2025b.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 61092–61107, 2024.

A Theoretical Analysis

A.1 Properties of Causal-order-constrained Attribution

Fix a finite action set Ω, an acyclic precedence relation $\prec$, and a deterministic replay game

$$v(S) = \hat{R}_T(\tau^S) - \hat{R}_T(\tau^\emptyset). \tag{11}$$

The event endpoint, reference, and replay randomness are held fixed. Thus $v(\emptyset) = 0$; verified full replay gives $v(\Omega) = R_T(\tau)$. We write $\phi_v^\prec$ below to expose the dependence suppressed in the main text. The following properties apply to this defined attribution; we do not claim uniqueness from the four ordinary Shapley axioms. The predecessor-constrained weighting follows the asymmetric Shapley framework (Frye et al., 2020).

$$\begin{aligned} \phi_v^\prec(a) &= \frac{1}{|\Pi_\prec(\Omega)|} \sum_{\pi \in \Pi_\prec(\Omega)} \Delta_a v(\pi), \\ \Delta_a v(\pi) &= v(\mathrm{Pre}_a(\pi) \cup \{a\}) - v(\mathrm{Pre}_a(\pi)). \end{aligned} \tag{12}$$

Here $\mathrm{Pre}_a(\pi)$ denotes the prefix before action instance a. Every legal prefix is closed under the specified predecessors. Empty precedence recovers the ordinary permutation average.

Property 1 (Efficiency). The attribution sums to the reference-relative endpoint risk:

$$\sum_{a \in \Omega} \phi_v^\prec(a) = v(\Omega) - v(\emptyset) = R_T(\tau). \tag{13}$$

Proof. For a legal order π, let S_j contain its first j actions. Then

$$\sum_{j=1}^{n} \big[v(S_j) - v(S_{j-1}) \big] = v(\Omega) - v(\emptyset). \tag{14}$$

Averaging this telescoping identity proves the result. It also holds for any finite average of complete sampled orders, irrespective of dependence or mixing, up to numerical replay error. □

Property 2 (Symmetry under constraint-preserving relabeling). Let g be a bijection on Ω such that $v(gS) = v(S)$ for all $S \subseteq \Omega$ and $a \prec b$ if and only if $g(a) \prec g(b)$. Then $\phi_v^\prec(g(a)) = \phi_v^\prec(a)$.

Proof. The map $\pi \mapsto g\pi$ is a bijection of $\Pi_\prec(\Omega)$, and $\mathrm{Pre}_{g(a)}(g\pi) = g(\mathrm{Pre}_a(\pi))$. Game invariance makes the corresponding marginal contributions equal. Their uniform averages are therefore equal. □

Ordinary symmetry without the precedence condition does not hold in general. For example, take $\Omega = \{a, b\}$, $a \prec b$, and $v(\{a, b\}) = 1$ with all other coalition values zero. The only admissible order yields $(\phi(a), \phi(b)) = (0, 1)$, although the game alone is symmetric. Ordinary Shapley yields $(1/2, 1/2)$.

Property 3 (Nullity). If $v(S \cup \{a\}) = v(S)$ for every $S \subseteq \Omega \setminus \{a\}$, then $\phi_v^\prec(a) = 0$.

Proof. Every marginal term $\Delta_a v(\pi)$ in Eq. 12 is zero, hence so is its average. The same conclusion holds if the zero-marginal condition is restricted to the admissible prefixes encountered before a. □

A negative contribution is not nullity: it denotes a risk-reducing contrast in the replay game.

Property 4 (Linearity for a fixed precedence relation). For games v_1, v_2 on the same Ω and the same $\prec$, and scalars α, β,

$$\phi_{\alpha v_1 + \beta v_2}^\prec(a) = \alpha \phi_{v_1}^\prec(a) + \beta \phi_{v_2}^\prec(a). \tag{15}$$

Proof. For each legal order, $\Delta_a(\alpha v_1 + \beta v_2)(\pi) = \alpha \Delta_a v_1(\pi) + \beta \Delta_a v_2(\pi)$. Averaging over the same order distribution proves the identity. Changing the graph between games would change that distribution and is outside this statement. □

Algorithm 1 Causal-order-constrained Shapley Attribution by Replay

Require: Action instances Ω, fixed precedence $\prec$, budget M, replay game v
Ensure: Approximate scores $\widehat{\phi}(a)$ for all $a \in \Omega$
Initialize the admissible-order sampler and $h(a) \leftarrow 0$ for all a
Evaluate and cache $v(\emptyset)$
for $m = 1$ to M **do**
  Draw π_m by block shuffling or the retained MCMC step
  Assert $\pi_m \in \Pi_{\prec}(\Omega)$; set $S \leftarrow \emptyset$, $u \leftarrow v(\emptyset)$
  for $j = 1$ to $|\Omega|$ **do**
    $a \leftarrow \pi_m[j]$; $S \leftarrow S \cup \{a\}$
    $w \leftarrow v(S)$ {Replay in physical chronology, or read cache}
    $h(a) \leftarrow h(a) + w - u$; $u \leftarrow w$
  end for
end for
return $\widehat{\phi}(a) = h(a)/M$ for all a

A.2 Estimating Causal-order-constrained Shapley Values

Enumerating all legal permutations is infeasible for full trajectories. We instead approximate Eq. 12 by replaying its sampled prefixes. The risk game is identical for our method and the Ordinary Shapley baseline; only the distribution of coalition inclusion orders differs.

A.2.1 Admissible-order Sampling

For EconAgent, the specified price-mediated dependencies give complete chronological blocks. Independently shuffling each block and concatenating blocks in chronological order samples uniformly from its legal extensions. Ordinary Shapley independently shuffles all n players.

For the general partial graphs in TwinMarket and SocialNetwork, we maintain four independently seeded chains. Starting from randomized topological orders, a transition stays in place with probability $1/2$; otherwise it chooses one of the $n-1$ adjacent pairs uniformly and swaps it only when the pair is incomparable. Each chain uses $10n^2$ burn-in transitions, followed by $2n^2$ transitions between its retained draws. We collect draws round-robin across chains and check every retained order against all graph edges.

The transition is symmetric on legal orders; adjacent incomparable swaps connect the linear extensions, and the self-loop makes the chain aperiodic. Thus its stationary distribution is uniform. This stationary-target argument supplies no finite-budget mixing bound. Randomized topological initialization is not itself a uniform sampler, and finite MCMC estimates are not claimed to be independent or unbiased. Position diagnostics and independently seeded attribution estimates are empirical checks rather than proofs of convergence.

A.2.2 Sample-average Estimator

For retained orders $\pi_1, \ldots, \pi_M$, define

$$\widehat{\phi}(a) = \frac{1}{M} \sum_{m=1}^{M} \Delta_a v(\pi_m). \tag{16}$$

This is an exact definition of the estimator, approximating the target in Eq. 12. The graph and the replay protocol are frozen for the run. Risk values may be cached by coalition without changing the estimator. Algorithm 1 describes both sampling cases.

A.2.3 Computational Cost

With $n = |\Omega|$, the algorithm requires at most $1 + Mn$ replay evaluations; caching can reduce this count. If one replay costs C_{sim}, total time is

$$O\big(C_{\mathrm{graph}} + C_{\mathrm{order}} + (1 + Mn)C_{\mathrm{sim}}\big), \tag{17}$$

where graph construction costs C_{graph} and order generation and validation cost C_{order}. Block shuffling costs $O(Mn)$ before validation. For four MCMC chains with burn-in B and retained-draw spacing L, transition overhead is $O(4B + ML)$ with constant-time edge lookup. Checking every order costs $O(M(n + |E|))$ for graph edge set E. Thus linear growth in replay calls does not imply linear total runtime for the chosen $B, L \propto n^2$.

A.3 Causal Constraints: Scope and Assumptions

Our use of causal knowledge has two distinct components: the supplied precedence relation controls admissible attribution contexts, while the replay game defines the risk contrast. Specifying one does not determine the other.

What the attribution measures. Equation 2 allocates reference-relative terminal risk over predecessor-respecting action coalitions. It adopts the ordering mechanism of asymmetric Shapley values (Frye et al., 2020), using a simulator replay game rather than their conditional feature-expectation game. Approaches based on interventional expectations define a different value function (Heskes et al., 2020). With downstream LLM choices held at their recorded values, the scores quantify action contributions in the replay game. This analysis can identify candidate targets for prevention, whose intervention effects require evaluation through policy rollouts.

Interactions and allocation depend on the graph. Nonadditive coalition values retain synergistic and redundant replay interactions, but Eq. 2 averages only admissible contexts. Specified predecessors enter before descendants; unspecified relations remain unrestricted. This changes how joint contributions are allocated, without removing interactions from v. Efficiency still holds. Accordingly, a large G_{ag} describes concentrated attribution magnitudes under the chosen graph and reference.

Counterfactual replay assumptions. Kept actions are recorded decisions applied to evolving simulator states; omitted actions use the fixed reference. Orders or interactions can consequently have different outcomes from the factual trajectory, including nonexecution. Replay follows the simulator's feasibility and matching rules and does not assume every recorded choice remains behaviorally plausible in every modified state. Predecessor closure excludes specified ordering violations, but cannot guarantee that all replay states lie on the original data distribution. Full-coalition endpoint agreement and repeated-coalition determinism are necessary implementation checks, not proofs of unrestricted causal validity.

Source and coverage of causal constraints. We freeze graph rules from simulator mechanisms and recorded dependencies before estimating attributions. To construct the graph, trace each action through the state or information it updates to later decisions that read it, retaining the mechanism or recorded dependency supporting each edge. Temporal precedence alone is insufficient. In EconAgent, labor and consumption affect aggregate supply, demand, and prices read by subsequent agents; these cross-step constraints yield chronological blocks. TwinMarket uses each agent's persistent history and submitted-stock price dependencies on later recorded stock decisions. SocialNetwork uses own-state continuity and recorded post/interaction feedback into next-step state. These encode possible structural influence, not a claim that every edge has a nonzero realized effect. TwinMarket forum channels and within-step dependencies inside SocialNetwork's composite actions are not covered. The graph is partial and remains fixed during replay.

A useful implication of complete chronological blocks is

$$\phi_t^{\text{tm}} = v(\Omega_{\leq t}) - v(\Omega_{<t}), \tag{18}$$

where $\Omega_{\leq t}$ contains all actions through t. Within-block terms telescope for every legal order, so these temporal sums need no permutation averaging. Both values evaluate risk at the fixed terminal horizon T, with later actions replaced by references; they are not observed risks at time t. This identity does not hold for arbitrary partial graphs.

Aggregation assumptions. Each agent contributes one action record per step through the fixed horizon T, including hold/no-op records. Time and agent aggregates partition these player records. A TwinMarket action contains the day's decisions over multiple stocks; a SocialNetwork action contains posting and all interactions in the step. Behavior aggregation therefore uses weights $w_{i,t,k}$, not necessarily a single category indicator. The weights sum to one for represented actions; records omitted from behavior categories must have individually verified zero attribution. These conditions

preserve the signed sum $R_T(\tau)$ at all three aggregation levels. Heterogeneous agent lifecycles are outside the present evaluation.

B SIMULATION SCENARIOS

Trajectory selection and sample counts. We select extreme-event trajectories using the scenario-specific criteria below. The corpus contains 142 trajectories: 39 from EconAgent ($N = 10$, $T \in [15, 50]$), 25 from TwinMarket ($N = 20$, $T \in [5, 42]$), and 78 from SocialNetwork ($N = 20$, $T \in [10, 30]$). Here T is the selected event endpoint, and each trajectory contains 100–840 agent–time actions. In the model order GPT, Llama, Claude, Qwen, and DeepSeek, the trajectory counts are $(8, 10, 9, 5, 7)$ for EconAgent, $(5, 2, 5, 8, 5)$ for TwinMarket, and $(15, 12, 10, 11, 30)$ for SocialNetwork. Table 2 uses these counts for the four agent and behavior metrics. For L_{tm}, the EconAgent GPT and Claude groups have 7 and 8 valid trajectories, respectively, because GPT seed 56 and Claude seed 44 have nonpositive terminal excess risk; the other counts are unchanged. The same two cases are excluded from Figure 5, leaving 140 trajectories across 15 scenario–model settings.

The trajectories are grouped by the archived model labels. Exact generation-model versions and decoding parameters were not retained in the available records.

Randomness control. EconAgent recovers the original initialization seed from the run identifier, retains the original episode length, and restores the archived reset and per-step NumPy random states. TwinMarket and SocialNetwork reset replay randomness to seed 42 for each coalition. This defines a deterministic, common-seed replay protocol; it does not guarantee matching individual exogenous draws after interventions change the sequence of random-number consumption. Permutation seeds are separate from replay seeds, and the 20 convergence runs use distinct permutation seeds on the same fixed trajectory. Input and implementation hashes accompany each estimate. Dataset selection and effective sample counts are reported separately from Monte Carlo repetitions.

B.1 MACROECONOMIC SIMULATION: ECONAGENT

The **EconAgent** environment simulates a macroeconomic ecosystem based on a one-step economy scenario. In this system, multiple agents engage in economic activities through labor supply and consumption decisions, while the system dynamically computes aggregate indicators such as price levels, wages, and inflation rates. Each agent's action space comprises two primary dimensions: *SimpleLabor*, where agents make a binary choice to supply labor (work) or not, and *SimpleConsumption*, where agents select a specific consumption rate.

Each agent's action space consists of two dimensions: *SimpleLabor*, a binary decision indicating whether the agent supplies labor, and *SimpleConsumption*, representing the chosen consumption ratio. To analyze behavior risk concentration, agent behaviors are categorized into eight discrete action types, defined by the Cartesian product of labor participation (work or not work) and four consumption intervals: $[0, 25\%)$, $[25, 50\%)$, $[50, 75\%)$, and $[75, 100\%]$.

In Figure 5(a), actions are indexed according to this combination of labor participation and consumption intensity. Specifically, actions 1–4 correspond to agents choosing not to work, with consumption ratios falling into the four increasing intervals listed above, while actions 5–8 correspond to agents choosing to work, with the same consumption intervals ordered from low to high.

For a coalition replay, actions outside the coalition are replaced by a fixed reference: supplying labor and consuming a proportion 0.8 (encoded consumption value 40). Actions inside the coalition retain their recorded decisions. This reference defines the contrast; it need not minimize risk. Each recorded action has unit weight in its work/consumption category.

System risk is quantified using the conditional variance of inflation forecast errors based on a naive forecast approach. For recorded prices P_t, define log-price inflation $\pi_t = \ln P_t - \ln P_{t-1}$. The expected inflation rate is defined as $E_{t-1}[\pi_t] = \pi_{t-1}$, yielding a forecast error of $e_t = \pi_t - E_{t-1}[\pi_t]$. The risk indicator $\hat{R}_t(\tau)$ is computed using an Exponential Weighted Moving Average (EWMA) model:

$$\hat{R}_t(\tau) = \lambda \cdot \hat{R}_{t-1}(\tau) + (1 - \lambda) \cdot e_{t-1}^2, \tag{19}$$

where $\lambda \in [0, 1]$ is the EWMA decay factor. The first available squared forecast error initializes the recursion; earlier undefined positions return zero in replay. We set $\lambda = 0.94$, consistent with the RiskMetrics standard. This metric captures the volatility of inflation dynamics and serves as a proxy for systemic macroeconomic risk.

We freeze T at the historical maximum of the raw inflation-risk indicator. The archived filter excludes trajectories whose maximum occurs at step 14 or earlier; this is a timing filter, not an expert-annotated risk-amplitude threshold. Attribution uses the signed contrast $R_T(\tau) = \hat{R}_T(\tau) - \hat{R}_T(\tau^{\emptyset})$ at this fixed endpoint, including cases where it is nonpositive.

B.2 Financial Market Simulation: TwinMarket

TwinMarket simulates a financial trading system in which agents manage portfolios across ten distinct industry-level stock indices. At each timestep, an agent makes trading decisions independently for each index. For a given index, the agent may choose to *buy* (by specifying order quantity and limit price), *sell* (by specifying order quantity and limit price), or *hold* (taking no action).

For behavior aggregation, categories identify the stock indices on which orders were submitted. The ten tradable indices, ordered consistently throughout the experiments, in Figure 5(b) are: `TLEI`, `MEI`, `CPEI`, `IEEI`, `REEI`, `TSEI`, `CGEI`, `TTEI`, `EREI`, and `FSEI`. A daily action may submit orders for several indices. Its attribution is split equally across the distinct indices with at least one submitted order, including unfilled orders; multiple orders in the same index do not increase its weight. A no-order action receives no category weight and is checked individually to have zero contribution.

For coalition replays, an omitted action is replaced by *hold*, meaning no trading orders are submitted for that agent and day. Retained actions reuse their recorded orders. This reference defines the attribution contrast without claiming to be a globally minimum-risk policy. Replays start on the first recorded trading day, with initial stock prices given by the most recent prior closes, and run through the fixed event date.

The risk modeling framework in TwinMarket mirrors that of EconAgent, employing an EWMA-based volatility measure. System risk is derived from a composite market price P_t: within-stock transaction VWAPs are weighted by the stock-profile weights, renormalized over stocks traded that day. Leading days without trades are omitted from the price series; later days without trades carry forward the previous composite market price. The price change rate is defined as the log return:

$$\pi_t = \ln P_t - \ln P_{t-1}. \tag{20}$$

Using a naive forecast $E_{t-1}[\pi_t] = \pi_{t-1}$, risk follows Eq. 19 with $\lambda = 0.94$. The recursion is initialized by the first squared forecast error; fewer than three prices give risk zero. These conventions reproduce the archived risk calculator, including no-trade coalitions.

For each retained archive trajectory, T is the first date attaining the maximum raw-risk value in its current market metrics. The full trading calendar through T is replayed, including dates omitted by older attribution summaries. We do not infer an expert-annotated amplitude threshold from these files. Attribution uses $R_T(\tau) = \hat{R}_T(\tau) - \hat{R}_T(\tau^{\emptyset})$, without reselecting the event under each coalition.

B.3 Social Dynamics Simulation: SocialNetwork

SocialNetwork models the dynamics of opinion polarization within a social media network, where agents interact through content creation and engagement. Each agent's action space consists of two main dimensions: *posting*, a binary decision indicating whether the agent publishes content, and *interaction*, where agents respond to each viewed post by selecting one of three options: *like*, *dislike*, or *not interact*. Consequently, for each viewed post, the agent has six possible action types corresponding to the combination of posting and interaction decisions.

For behavior analysis, agent actions are categorized into six discrete types corresponding to the Cartesian product of posting and interaction decisions: *not post–like*, *not post–not interact*, *not post–dislike*, *post–like*, *post–not interact*, and *post–dislike*. When a step has multiple recorded interactions, its attribution is divided equally among those interactions and assigned by posting status and interaction type. A posting-only step maps to *post–not interact*. A step with neither posting nor interactions has no category weight and is checked to have zero attribution.

In Figure 5(c), actions are indexed according to this combination of posting and interaction decisions. Specifically, actions 1–3 correspond to agents choosing not to post, with interactions *like*, *not interact*, and *dislike*, respectively. Actions 4–6 correspond to agents choosing to post, with the same three interaction options in the order *like*, *not interact*, and *dislike*.

In each coalition replay, only actions outside the coalition are replaced by the no-post/no-interaction reference. Retained players reuse the complete logged composite action. Initial beliefs come from the recorded initial state; the endpoint after the first T action steps corresponds to action indices $0, \ldots, T-1$ in the implementation.

B.3.1 Belief Update Mechanism

Agent beliefs evolve through two primary mechanisms: *interaction-based updates* and *feedback-based updates*. The update equations below describe individual update stages: each uses the current belief and is followed by clipping to $[-1, 1]$. Interactions are applied sequentially, followed by feedback; risk is computed from the beliefs after all updates in the step.

Interaction-based updates occur when an agent engages with content posted by others. When an agent i with belief b_i^t interacts with a post containing belief value b_p, the belief update depends on the interaction type:

$$b_i^{t+1} = \begin{cases} b_i^t + \delta \cdot s_i^t \cdot |b_p - b_i^t| & \text{if like and } b_p > b_i^t \\ b_i^t - \delta \cdot s_i^t \cdot |b_p - b_i^t| & \text{if like and } b_p \leq b_i^t \\ b_i^t - \delta \cdot s_i^t \cdot |b_p - b_i^t| & \text{if dislike and } b_p > b_i^t \\ b_i^t + \delta \cdot s_i^t \cdot |b_p - b_i^t| & \text{if dislike and } b_p \leq b_i^t \end{cases} \tag{21}$$

where δ is the update magnitude parameter, and s_i^t is the belief update sensitivity of agent i at time t, defined as:

$$s_i^t = s_{\text{base}} \cdot (1 - \alpha \cdot |b_i^t|), \tag{22}$$

with s_{base} being the base sensitivity and α a sensitivity coefficient. This formulation ensures that agents with more extreme beliefs (higher $|b_i^t|$) exhibit reduced sensitivity to new information, reflecting the phenomenon of confirmation bias and belief rigidity.

Feedback-based updates occur when an agent receives engagement on their own posts. After agent i publishes a post at timestep t, if the post receives at least one view ($V_i^t > 0$), the belief is updated based on the net engagement received:

$$b_i^{t+1} = b_i^t \cdot \left(1 + \frac{L_i^t - D_i^t}{V_i^t} \cdot \eta\right), \tag{23}$$

where L_i^t, D_i^t, and V_i^t denote the number of likes, dislikes, and total views received by agent i's post at timestep t, respectively, and η is the reinforcement coefficient (re-denoted from ρ to avoid a clash with the extreme-event threshold in the main paper). This mechanism models social reinforcement, where positive feedback (net likes) strengthens the agent's current belief, while negative feedback (net dislikes) weakens it. If no views are received ($V_i^t = 0$), the belief remains unchanged.

Additionally, agent behavioral parameters—posting preference and interaction preference—are dynamically updated based on the agent's current belief magnitude:

$$p_i^t = p_{\text{base}} + \beta \cdot |b_i^t|, \tag{24}$$

$$q_i^t = q_{\text{base}} + \gamma \cdot |b_i^t|, \tag{25}$$

where p_i^t and q_i^t represent the posting and interaction preferences, respectively, p_{base} and q_{base} are base preference values, and β and γ are preference coefficients. This design captures the observation that agents with more extreme beliefs tend to be more active in both content creation and engagement.

System risk is quantified as the level of belief polarization in the network, measured by the variance of agent beliefs at time t:

$$\hat{R}_t(\tau) = \frac{1}{N} \sum_{i=1}^{N} (b_i^t - \bar{b}^t)^2, \tag{26}$$

where b_i^t represents the belief value of agent i at time t, $\bar{b}^t$ is the mean belief across the population, and N is the total number of agents. The archived filter requires the raw-risk maximum to occur at step 10 or later and to be at least 2.5 times the positive initial risk. We freeze T at the recorded raw-risk maximum and attribute the signed excess $R_T(\tau) = \hat{R}_T(\tau) - \hat{R}_T(\tau^{\emptyset})$ there; the selection threshold is distinct from this reference-relative contrast.

C Implementation Details of Baseline Methods

C.1 Failure Taxonomy (FT)

Cemri et al. (2025) develop a taxonomy of 14 failure modes covering specification and system design, inter-agent misalignment, and task verification and termination. We adapt taxonomy-guided prompting to dense action-risk scoring in the controlled games. The prompt supplies concise paraphrases of the 14 categories and two author-created examples. It asks the model to use a category only when supported by observable evidence and to assign one score per action, rather than category scores or invented failure labels.

C.2 Failure Attribution (FA)

Following the trajectory-based attribution approach of Zhang et al. (2025b), the FA prompt inspects the action sequence and dependencies to score every action for the experiment's specified risk target, rather than selecting a single earliest decisive error. Both FT and FA are task-specific dense-scoring adaptations of failure-analysis approaches.

Matched inputs and outputs. For the extension diagnostics in Appendix D.9, both methods receive every action, all prerequisite edges, the outcome equation and coefficients, and the complete coalition-value lookup. Each returns one score in $[0, 1]$ per action; the shared instruction assigns zero to actions judged irrelevant or protective and requests no score normalization. Scores are normalized only for evaluation. The methods share the model and decoding settings specified in Appendix D.9; only the taxonomy context and method-specific instruction differ. Neither prompt instructs the model to imitate ASV or optimize the reported diagnostics. For the intervention benchmark in Appendix D.10, the shared target is instead the stage-local risk increment; the complete lookup and graph are supplied, while coefficients and intervention labels are withheld. The taxonomy and sequence-based prompting distinction is retained.

C.3 Shapley Value (SV) and Leave-One-Out (LOO)

Both baselines use the same action set Ω, fixed reference actions, replay randomness, and characteristic function v as the constrained method. Let $n = |\Omega|$. **SV** (Shapley, 1953) removes the precedence restriction and assigns

$$\phi^{\mathrm{Ord}}(a) = \sum_{S \subseteq \Omega \setminus \{a\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \big[v(S \cup \{a\}) - v(S)\big]. \tag{27}$$

We estimate it with uniformly shuffled permutations for the simulation trajectories and evaluate the subset sum exactly for the controlled games. It is a baseline with a different allocation rule, not an earlier version of our method.

LOO (Covert et al., 2021) uses single-action removal attribution and assigns

$$\mathrm{LOO}(a_{i,t}) = v(\Omega) - v(\Omega \setminus \{a_{i,t}\}) = \widehat{R}_T(\tau) - \widehat{R}_T(\tau^{\Omega \setminus \{a_{i,t}\}}). \tag{28}$$

The all-reference term in v cancels. This score replaces one recorded action while retaining all others and evaluates the risk at the same endpoint. It requires one replay per action after the shared factual evaluation. LOO does not average over coalition contexts and need not sum to the full-minus-reference risk when actions interact.

AgenTracer (Zhang et al., 2025a) trains a failure-localization model using curated counterfactual replay and fault-injection data. We do not train or evaluate that model here; the direct single-action replay comparator is labeled LOO.

Table 5: Scale-aware MRSD (%) over 20 independent runs on one fixed trajectory per system; budgets are nested prefixes of the same runs.

Sample size M	EconAgent	TwinMarket	SocialNetwork
$M = 10$	25.48 %	67.07 %	43.68 %
$M = 100$	11.16 %	29.91 %	17.07 %
$M = 300$	7.28 %	19.99 %	10.65 %
$M = 1,000$	4.61 %	13.01 %	6.08 %

D Additional Experiments and Analyses

D.1 Exact Checks on Enumerable Games

Table 1 evaluates the sampler on four fully specified six-player games with player labels $0, \ldots, 5$. The common value function is

$$v(S) = \sum_{i \in S} w_i + 1.5\,\mathbf{1}[\{0,2\} \subseteq S] + 0.7\,\mathbf{1}[S \cap \{1,3\} \neq \emptyset] - 0.6\,\mathbf{1}[\{2,4,5\} \subseteq S] + 0.2|S|^2, \tag{29}$$

with $w = (-0.5, 0.75, 1, 0.25, -0.25, 0.5)$. The cases use: (i) the chain $0 \prec 1 \prec \cdots \prec 5$; (ii) edges $(0,2), (1,2), (1,3), (2,4), (3,5)$; (iii) no edges; and (iv) all edges from $\{0,1,2\}$ to $\{3,4,5\}$. Enumerating legal permutations gives respectively 1, 16, 720, and 36 exact marginal-contribution vectors to average. We apply the sampler from Appendix A.2, including its MCMC settings for the first two cases, using 10 independent seeds and nested budgets. The single-order chain is deterministic; its displayed unit similarity is a structural consequence, not empirical evidence of mixing. At $M = 1{,}000$, mean relative ℓ_2 errors are respectively below 10^{-12}, 0.496%, 1.334%, and 1.117%.

D.2 Monte Carlo Convergence Study (MRSD)

For one fixed trajectory per scenario, our protocol uses $K = 20$ independently seeded attribution runs with a maximum budget of $M = 1{,}000$. Estimates at $M \in \{10, 100, 300, 1{,}000\}$ are nested prefixes within each run, rather than independent experiments at each budget. Let $d_{i,t} = K^{-1} \sum_k |\widehat{\phi}^{(k)}(a_{i,t})|$. We report scale-aware mean relative standard deviation:

$$\mathrm{MRSD} = \frac{100\%}{NT} \sum_{i,t} \frac{\mathrm{std}_k(\widehat{\phi}^{(k)}(a_{i,t}))}{d_{i,t} + \varepsilon}, \qquad \varepsilon = \frac{10^{-3}}{NT} \sum_{i,t} d_{i,t}, \tag{30}$$

using sample standard deviation across runs. The stabilizer scales with attribution magnitude, avoiding an absolute constant that can dominate small risk units. If all scores are zero, we define MRSD as zero and flag the degenerate case. We also inspect absolute variation, near-zero scores ($d_{i,t} \leq \varepsilon$), pairwise cosine similarity, Spearman correlation, and signed Top-10 overlap. Pairwise comparisons share runs and are not treated as independent samples.

At $M = 1{,}000$, the completed EconAgent and SocialNetwork cases have MRSD 4.61% and 6.08%, respectively. Across all pairs of their 20 runs, minimum cosine similarities are 0.999872 and 0.999540, minimum Spearman correlations are 0.998699 and 0.999128, and minimum Top-10 overlaps are 100% and 90%. These cases meet the prespecified ranking diagnostics (cosine ≥ 0.99, Spearman ≥ 0.95, Top-10 overlap ≥ 0.8); they do not support a blanket MRSD-below-5% claim.

TwinMarket has MRSD 13.01%, minimum pairwise cosine 0.986578, minimum Spearman 0.899485, and minimum Top-10 overlap 90%. It fails the prespecified cosine and rank-correlation targets at $M = 1{,}000$. The relative ℓ_2 standard deviation is 9.17%; 344 of 680 players have mean absolute score at or below the scale-aware stabilizer. Maximum position $\widehat{R}$ across the 20 runs is 1.027, which does not by itself establish sufficient attribution precision. Thus $M = 1{,}000$ is not uniformly adequate across these three cases.

Separate financial precision check. We added 20 independent $M = 1{,}000$ runs on the same fixed TwinMarket trajectory, giving 40 in total, and used the prespecified disjoint groups of run

Table 6: Reference sensitivity on five fixed TwinMarket trajectories: mean ± sample SD across trajectories. No-op is an identity check at $M = 10$; the other rows use $M = 1{,}000$.

Reference	Cosine similarity	Spearman correlation
Noise	0.027 ± 0.085	-0.009 ± 0.051
Historical resampling	0.111 ± 0.138	0.072 ± 0.089
Historical mean	0.098 ± 0.157	0.094 ± 0.055
No-op (identity)	1.000 ± 0.000	1.000 ± 0.000

indices 0–7, 8–15, 16–23, 24–31, and 32–39. Each group averages eight action-score vectors, for an effective budget of $8{,}000$ permutations. Across the five groups, minimum cosine similarity is 0.998802, minimum Spearman correlation is 0.960844, and minimum Top-10 overlap is 1.0; all three original thresholds are met. This supports an eight-run averaging budget for the financial main experiment. There are five independent groups, not five independent trajectories; their ten pairwise comparisons share groups. Nonlinear metrics must be recomputed from each averaged action-score vector. Table 5 retains the original 20-run, maximum-$1{,}000$ protocol and its financial failure result.

D.3 Sensitivity to Reference Actions

References specify the contrast being attributed (Lundberg & Lee, 2017; Sundararajan et al., 2017). We use five prespecified TwinMarket/GPT-group trajectories (runs 42–46; $N = 20$), holding each factual horizon, precedence graph, and replay seed fixed. Each non-default comparison uses $M = 1{,}000$ per vector. Histories below use the first submitted sub-order per user–stock record over the fixed horizon; they define post-hoc references, not online policies.

Noise preserves an active order's side and independently perturbs its first price and quantity with Gaussian standard deviations $\max(10^{-12}, 0.1\,\mathrm{SD})$, using that user's stock-specific history; inactive records remain hold. **Historical resampling** samples a price–quantity–side tuple from that history. **Historical mean** uses its mean price and quantity and majority side (ties: first observed). Prices are floored at 10^{-8}, quantities at zero; zero quantity or missing history yields hold. Random references are generated once with seed 20260926 and frozen across coalitions. **No-op** equals the default hold reference and is checked only against the matching $M = 10$ prefix. We compare signed vectors without rescaling.

The alternative references yield substantially different rankings. Noise produces negative factual-minus-reference risk in two trajectories; we retain these signed contrasts rather than treating every reference as safer. Changing the reference changes the attribution question. These estimates combine reference sensitivity with finite-sampling error: the $M = 1{,}000$ financial precision limitation in Appendix D.2 still applies.

D.4 Protocol for the Parameter Sensitivity Analysis

Section 4.4 and Table 4 report the sensitivity results. Here we specify the two protocols. First, an additional subset rule keeps EconAgent trajectories with terminal excess risk $v(\Omega) \geq \rho$. This is a supplementary selection rule, not a reconstruction of the original event definition. We freeze $\rho_0 = 1.654776 \times 10^{-4}$, the median of the 37 positive excess risks in the full 39-trajectory economic corpus. Thresholds $0.8\rho_0$, ρ_0, and $1.2\rho_0$ retain 25, 19, and 15 trajectories. Each retained trajectory keeps its endpoint, action set, reference, graph, and replay randomness. Its attribution therefore stays unchanged, while averages over different subsets can change. Table 4 reports trajectory-weighted means and sample standard deviations of causal-order-constrained metrics within each complete subset. All five metrics are defined for every retained trajectory.

Second, changing q recomputes only the first strict crossing used by L_{tm}, with the signed action contributions fixed. The lower panel of Table 4 uses one economic trajectory ($N = 10$, $T = 28$, $M = 1{,}000$); the other four metrics are invariant by definition. This trajectory is outside all three threshold subsets above, so the two panels are not matched-cohort comparisons.

Table 7: HHI-selected EconAgent events: mean ± sample SD, with n valid trajectories for each of the five metrics.

Model	n	L_{tm}	G_{ag}	C_{ag}	Z_{ag}	G_{be}
Llama	4	0.013 ± 0.025	0.791 ± 0.009	0.996 ± 0.004	0.639 ± 0.051	0.745 ± 0.030
Qwen	3	0.068 ± 0.072	0.755 ± 0.020	0.973 ± 0.024	0.671 ± 0.110	0.747 ± 0.019

For the fixed trajectory, the first strict crossings occur at steps 11, 14, and 15 for $q = 0.80$, 0.90, and 0.95, respectively; $L_{\text{tm}} = (T - t_q)/T$. The agent and behavior metrics remain unchanged because the action attributions are fixed.

D.5 Alternative Risk Metric (Herfindahl–Hirschman Index)

We define wealth concentration as $H_t = \sum_i (w_{i,t}^+ / \sum_j w_{j,t}^+)^2$, where $w_{i,t}^+ = \max(w_{i,t}, 0)$ is nonnegative Coin endowment; $H_t = 0$ when total wealth is at most 10^{-18}. The pooled 95th percentile over 3,950 recorded timesteps from 79 available trajectories is 0.618624. Eligible events have a peak at least this threshold and a latest peak time $T \geq 15$. Keeping the original Llama/Qwen comparison scope yields four Llama and three Qwen trajectories ($N = 10$), all included; no runs are added to equalize sample sizes.

Each event uses its HHI-selected endpoint and the same EconAgent replay initialization, recorded randomness, reference actions, and chronological-block constraints. We estimate contributions with $M = 1{,}000$ and report trajectory-level means and sample SD. HHI selects different events and horizons from the original risk definition, so this is not a paired comparison on the same events.

These selected events show concentrated agent and behavior contributions, while the mean 90% cumulative-contribution crossing occurs near the event endpoint. Mean C_{ag} is positive in both groups, relating absolute agent contributions to their temporal variability. These HHI estimates have not undergone a separate 20-run precision study.

D.6 Wall-clock Time and API Cost

We distinguish trajectory generation, attribution, and baseline evaluation. Generation uses the recorded LLM requests; its wall time and API cost depend on request count, tokens, concurrency, and retries. FT/FA judging incurs API costs, which are accounted for separately from simulator replay costs. Direct LOO uses simulator replays.

For $n = NT$ action players, let C_{sim} denote the cost of one *complete fixed-horizon replay*, consistently with Appendix A.2. The attribution computation costs

$$O(C_{\text{graph}} + C_{\text{order}} + (1 + MNT)C_{\text{sim}}) . \tag{31}$$

If a replay performs T global environment steps with average cost C_{step}, then $C_{\text{sim}} = O(TC_{\text{step}})$ and the replay component is $O((1+MNT)TC_{\text{step}})$. Step cost itself can depend on agent count and scenario dynamics. General-order MCMC also incurs burn-in, spacing, and order-validation costs; these must not be dropped from runtime estimates. Averaging eight independent $1{,}000$-permutation runs repeats their setup and burn-in, even though the total retained budget is $8{,}000$.

Per-run audits record requested workers and available CPU affinity. The current runner records wall time, summed worker CPU time, setup and order-generation time, and maximum per-worker resident memory. Earlier archived runs do not contain all timing components; missing timings cannot be inferred from their totals. These quantities have different scopes: worker CPU time covers order evaluation rather than all job processes, maximum worker memory is not total job memory, and overlapping timing components must not be added as if they were disjoint. Separate worker caches can repeat endpoint evaluations. Final runtime comparisons must state these boundaries, budget, horizon, and worker count; a single historical per-trajectory time is not a hardware-independent cost estimate.

Hardware and software. The revised replay experiments use CPU-only Linux container jobs with Python 3.10.12 and a shared dependency directory containing NumPy 2.2.5, SciPy 1.15.2, pandas

Table 8: Attribution costs: mean ± trajectory sample SD. t_{1000} is estimator-internal wall time per 1,000 permutations; C_{worker} sums evaluation-worker CPU hours over the full budget M. These are different resource measures.

Scenario	Method	n	M	t_{1000} (s)	C_{worker} (CPU h)
EconAgent	Causal-order	38	1000	1642.9 ± 1051.4	6.756 ± 4.333
EconAgent	Ordinary	38	1000	2998.3 ± 1931.3	12.290 ± 7.920
TwinMarket	Causal-order	24	8000	654.1 ± 471.7	19.683 ± 16.159
TwinMarket	Ordinary	24	8000	578.9 ± 434.9	19.166 ± 14.411
SocialNetwork	Causal-order	78	1000	328.0 ± 229.3	1.075 ± 0.899
SocialNetwork	Ordinary	78	1000	251.2 ± 180.7	1.040 ± 0.749

2.2.3, and PyYAML 6.0.2. The container image is `ml-base:22.04-pjlab-20251117`; frozen scripts and input hashes are retained with each estimate. Worker allocations vary across jobs and are recorded per run. This describes the replay environment, not the unverified environment or API charges of historical trajectory generation. Table 8 summarizes 140 matched trajectories with 15 evaluation workers per run. EconAgent/GPT seed 42 and TwinMarket/GPT seed 42 contain preflight reuse and are excluded from timing only; both remain in the scientific cohort. Both nonpositive-contrast cases remain in timing. Allocated workers do not imply identical hardware or contention.

For TwinMarket, t_{1000} averages the eight run wall times within each trajectory, while worker CPU hours are summed. Both summaries then weight trajectories equally; parallel run times are not added into elapsed time. Timings exclude queueing, imports before the estimator timer, caller checks, postprocessing, historical generation, and FT/FA.

D.7 Frequency-Normalized Behavioral Attribution

For the weights in Eq. 9, define the effective occurrence count and intensity as

$$c_k = \sum_{t=1}^{T} \sum_{i=1}^{N} w_{i,t,k}, \qquad \hat{\phi}_k^{\text{be}} = \frac{\phi_k^{\text{be}}}{c_k} \quad (c_k > 0). \tag{32}$$

Weights split a composite action across its represented categories, so c_k can be fractional. No-order/no-post-no-interaction records without a category assignment are included in the player game, but are excluded from these counts only after checking their individual attribution is zero. Zero-occurrence categories have undefined intensity and are excluded from $\hat{G}_{\text{be}}$. The un-normalized G_{be} retains the full category vocabulary; either Gini is undefined when its absolute contributions sum to zero.

Total category contributions sum to the signed excess risk, whereas per-occurrence scores do not. The latter are descriptive averages under the same attribution game; dividing by frequency does not identify an intervention on behavior frequency, establish intrinsic harmfulness, or guarantee that the Gini decreases. Changing the set of observed categories also affects comparisons between the two Ginis.

Table 9 reports the three fixed validation cases, with category weights and signed mass independently reconstructed from the recorded actions. Model-group mean per-occurrence Gini values for causal-order-constrained attribution range from 0.495–0.596 in EconAgent, 0.449–0.552 in TwinMarket, and 0.284–0.529 in SocialNetwork. The measure is defined for every evaluated trajectory.

D.8 Sensitivity to the Precedence Structure

We hold the recorded actions, reference actions, replay randomness, and terminal risk fixed, and vary only the precedence constraints used for attribution. This comparison examines different structural assumptions, rather than small perturbations of a fixed graph. The four specifications are the scenario-specific default graph, each agent's own action history without cross-agent edges, chronological blocks (all actions at an earlier step precede all actions at a later step), and the empty order (Ordinary Shapley). Chronological blocks coincide with the default closure in EconAgent and are

Table 9: Behavior concentration on one fixed validation trajectory per scenario, using the prespecified first $M = 1{,}000$ estimate. The normalized Gini uses only categories with positive effective occurrence count.

Scenario	Total G_{be}	Per-occurrence $\hat{G}_{\text{be}}$	Observed/all types
EconAgent	0.741	0.507	5/8
TwinMarket	0.461	0.451	10/10
SocialNetwork	0.589	0.370	6/6

therefore not counted as another estimate. These comparisons use one prespecified GPT-group trajectory per scenario: EconAgent run 42 ($N = 10, T = 28$), TwinMarket run 42 ($N = 20, T = 34$), and SocialNetwork run 42 ($N = 20, T = 15$).

For the own-history control, we sample uniform linear extensions exactly: shuffle a multiset containing one agent label per action, and consume the next chronological action of each selected agent. Every legal label sequence has the same number $\prod_i T_i!$ of preimages among permutations of the individually labeled tokens, where T_i is agent i's number of actions. This yields independent uniform draws. Chronological blocks use independent within-step shuffles; Ordinary Shapley uses unrestricted shuffles. The default TwinMarket and SocialNetwork graphs use the approximate adjacent-swap sampler described above. Each row uses the prespecified first $M = 1{,}000$ estimate; exact sampling of orders does not imply exact attribution values.

We report the fraction of unordered action pairs comparable in the graph's transitive closure and the Spearman correlation of signed action scores with the default estimate.

Table 10 quantifies how action rankings depend on the supplied structural assumptions. The TwinMarket comparison uses $M = 1{,}000$; its repeated-run precision results are reported in Appendix D.2.

D.9 Controlled Sequential-Game Evaluation

Game construction. We fix 96 games before evaluation: $n \in \{6, 8, 10\}$ actions, nominal forward-edge probability $p \in \{0, 0.2, 0.5, 0.8\}$, and eight seeds per combination. Indices are $0, \ldots, n-1$. The final action is isolated, has weight zero, and participates in no interaction. Each other action has an independent integer weight $w_i \in \{1, \ldots, 9\}$. Edges connect increasing indices among these non-null actions with probability p. For $p > 0$, each pair independently receives an interaction with probability 0.2, with equal probability of synergy $y_i y_j$ or redundancy $\max(y_i, y_j)$ and integer coefficient in $\{1, 2, 3\}$. For $p = 0$, interactions are omitted, giving 24 additive controls with known contributions w_i. Graphs sampled without edges in a nonzero-density regime are retained. Seeds are $20260926 + 10000n + 100d + r$, where $d \in \{0, 1, 2, 3\}$ indexes the density and $r \in \{0, \ldots, 7\}$ indexes the repeat.

For coalition S, let $x_i = \mathbf{1}\{i \in S\}$ and realize events in topological order as $y_i = x_i \prod_{j \in \text{pa}(i)} y_j$. The base value is

$$v_0(S) = \sum_i w_i y_i + \sum_{(i,j,c,f) \in \mathcal{H}} cf(y_i, y_j), \tag{33}$$

where $\mathcal{H}$ contains the sampled interaction terms. Thus $v_0(\emptyset) = 0$ and $E = v_0(\Omega) > 0$.

Extensions outside the admissible domain. Let $I_G(S)$ indicate that S omits a specified predecessor of one of its members. We evaluate six extensions

$$v_{\lambda,h}(S) = v_0(S) + \lambda E I_G(S) h(S), \tag{34}$$

with $\lambda \in \{-1, 1\}$ and $h(S) \in \{1, \sum_{i \in S}(i+1)/[n(n+1)/2], \binom{|S|}{2}/\binom{n}{2}\}$. Lookup values are rounded to six decimal places for all methods. Every admissible value and both endpoints remain unchanged. These extensions test dependence on values outside the specified domain; they are diagnostic extensions rather than new physical event mechanisms.

For FT/FA, extension prompts retained the base outcome equation and coefficients while supplying the extended coalition-value lookup, without an explicit precedence rule between these descriptions.

Table 10: Precedence sensitivity on three fixed trajectories ($M = 1{,}000$ per estimate). Comparable pairs are reported as fractions. TwinMarket's default estimate at this budget does not meet the repeated-run precision criterion.

Scenario	Specification	Comparable	Spearman
EconAgent	Default constraints	0.968	1.000
EconAgent	Own history only	0.097	0.947
EconAgent	Ordinary Shapley	0.000	0.897
TwinMarket	Default constraints	0.912	1.000
TwinMarket	Own history only	0.049	0.609
TwinMarket	Chronological blocks	0.972	0.915
TwinMarket	Ordinary Shapley	0.000	0.572
SocialNetwork	Default constraints	0.884	1.000
SocialNetwork	Own history only	0.047	0.982
SocialNetwork	Chronological blocks	0.936	0.998
SocialNetwork	Ordinary Shapley	0.000	0.845

Table 11: Controlled sequential-game evaluation: mean $\pm$ sample SD across games, in percent. D: inadmissible-extension sensitivity (72 games); Q: null-action share (96); A: additive recovery error (24); J: same-input variation (96). Lower is better for each diagnostic. FT/FA are dense-scoring adaptations using DeepSeek-V3.2.

Method	$D \downarrow$	$Q \downarrow$	$A \downarrow$	$J \downarrow$
SV (Shapley, 1953)	106.24 ± 27.69	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LOO (Covert et al., 2021)	172.06 ± 31.57	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
FT (Cemri et al., 2025)	64.57 ± 24.31	0.44 ± 1.48	7.78 ± 13.02	41.75 ± 38.90
FA (Zhang et al., 2025b)	72.99 ± 28.98	0.96 ± 5.31	15.72 ± 19.55	42.77 ± 41.09
ASV (ours) (Frye et al., 2020)	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Their D values therefore characterize sensitivity under this combined-input protocol, including possible responses to discrepancies between the two descriptions. The Q, A, and J diagnostics use the unextended game, where the descriptions agree.

Methods and repetitions. We enumerate all 2^n coalition values. SV uses exact subset weights; ASV uses exact prefix/suffix counts of legal orders; LOO uses full-minus-single-action differences. An independent rational-arithmetic implementation verifies all 2,016 exact score vectors. For FT and FA, we use `deepseek-ai/DeepSeek-V3.2` through SiliconFlow, temperature zero, thinking disabled, a 768-token output limit, and JSON output. Each method receives the complete game described in Appendix C. Two calls per method, game, and version give 2,688 valid responses after logged transport retries. We retain raw responses and validate model identity, unique action identifiers, numeric ranges, and completion status. Failed calls are logged and excluded from aggregation until successfully completed; they are never assigned zero scores.

Metrics. For any score vector s, define $N(s) = s/\|s\|_1$, with $N(0) = 0$. Write $b_m(v) = N(s_m(v))$ for an exact method and $b_m(v) = [N(s_m^{(1)}(v)) + N(s_m^{(2)}(v))]/2$ for FT/FA. With $\mathcal{V}$ denoting the six extensions, define

$$D_m = \max_{v \in \mathcal{V}} \|b_m(v) - b_m(v_0)\|_1, \qquad Q_m = |b_{m,n-1}(v_0)|, \tag{35}$$

$$A_m = \tfrac{1}{2}\|b_m(v_0) - N(w)\|_1, \qquad J_m = \|N(s_m^{(1)}(v_0)) - N(s_m^{(2)}(v_0))\|_1. \tag{36}$$

D is summarized over the 72 nominal nonzero-density games; Q and J over all 96 games; A over the 24 additive controls. For exact methods $J = 0$. All entries are multiplied by 100 and reported as means and sample standard deviations across games. Repeated API calls are not treated as additional games. D can reach 200%; same-input variation J contextualizes FT/FA sensitivity and is not subtracted from D.

ASV's invariance follows because every legal marginal uses admissible endpoints. The null-action and additive controls instead have answers specified independently by the outcome equation. These diagnostics distinguish structural invariance, irrelevant-action allocation, and recovery of known additive contributions; invariance alone is not a causal-effect accuracy label.

An independent Decimal-arithmetic implementation reconstructs the normalized profiles from the raw responses and verifies all 20 reported means and sample standard deviations, including their game counts. ASV has zero extension sensitivity; all three exact methods have zero null-action share and additive recovery error. FT and FA show nonzero errors and substantial same-input variation, as reported in Table 11.

D.10 Intervention-grounded Recovery of Stage-local Risk Contributions

Mechanism and intervention target. We construct 180 games with known causal prerequisites. For binary action requests x_i, realized events and accumulated risk satisfy

$$y_i = x_i \prod_{j \in \mathrm{pa}(i)} y_j, \qquad R(\mathbf{x}) = \sum_{i=0}^{n-1} \theta_i y_i, \tag{37}$$

evaluated in topological order, with an empty product equal to one. The factual requests are all one and reference requests are zero, giving $v(S) = R(\mathbf{1}_S)$ and $v(\emptyset) = 0$. Every coalition follows this same mechanism.

Let $\mathrm{An}(i)$ be the strict ancestors of action i. Set $x_j^{i,b} = 1$ for $j \in \mathrm{An}(i)$, $x_i^{i,b} = b$, and all other requests to zero. Paired simulator interventions measure

$$\delta_i = R\big(\mathrm{do}(\mathbf{x} = \mathbf{x}^{i,1})\big) - R\big(\mathrm{do}(\mathbf{x} = \mathbf{x}^{i,0})\big) = \theta_i. \tag{38}$$

This measures the risk added when the action executes after its ancestors, with other requests held at reference. The intervention labels are generated by scalar simulator runs independently of the attribution code. Blocking an action in the complete trajectory also suppresses dependent future events; LOO measures that distinct total removal effect.

Recovery property. For any legal prefix before action i, all its ancestors are present and none of its descendants are present. Adding i therefore activates only event i, yielding marginal contribution θ_i. Thus ASV recovers δ_i for every legal order in this mechanism class. To expose the baseline contrast, write $B_i = \mathrm{An}(i) \cup \{i\}$. The game is a sum of unanimity terms $v(S) = \sum_i \theta_i \mathbf{1}\{B_i \subseteq S\}$, so

$$\phi_j^{\mathrm{ASV}} = \theta_j, \qquad \phi_j^{\mathrm{SV}} = \sum_{i: j \in B_i} \frac{\theta_i}{|B_i|}, \qquad \phi_j^{\mathrm{LOO}} = \sum_{i: j \in B_i} \theta_i. \tag{39}$$

The exact-recovery guarantee is specific to the stated mechanism and stage-local intervention target.

Game design. We cross $n \in \{6, 8, 10\}$ with five graph families and twelve seeds, giving 144 dependent games and 36 independent controls. The final action is isolated with $\theta_{n-1} = 0$. The remaining coefficients are sampled without replacement from $\{0.010, 0.020, \ldots, 0.990\}$, independently of graph position. On indices $0, \ldots, n-2$, the graph families are: no edges; the chain $(i, i+1)$; the fork $(0, i)$; the join $(i, n-2)$; and a random DAG with each forward edge included with probability 0.350. Graphs sampled without edges are retained. Seeds are $20260926 + 10000n + 100f + r$, with family index $f \in \{0, \ldots, 4\}$ in the listed order and repeat $r \in \{0, \ldots, 11\}$. Two graph controls keep the game fixed while ASV uses either independently retained edges with probability 0.500, or the original graph with randomly permuted node labels. The latter remains acyclic but changes its correspondence with the mechanism.

Methods and model inputs. SV, LOO, and ASV are evaluated exactly on the same complete coalition-value table. FT and FA receive all action requests, the true graph, the event rule, and the complete consistent lookup. Risk coefficients and intervention labels are withheld. The shared instruction explicitly asks for the stage-local increment in Eq. 38; FT adds taxonomy context, whereas FA uses sequence-based prompting. Each returns one nonnegative relative score in $[0, 1]$ per action. We use DeepSeek-V3.2 through SiliconFlow, temperature zero, thinking disabled, a 768-token limit, and JSON output. Two calls per method and game yield 720 valid responses from 758 attempts, including transport and output-validation retries. Model identity, completion status, action IDs, and score ranges are validated. Actual zero-score predictions are retained.

Table 12: Intervention-grounded recovery: mean ± sample SD across games. The main-text table reports the corresponding means. Error and Top-1 are percentages; r_s is Spearman correlation.

Method	Error↓	Top-1↑	r_s ↑
Dependent games ($n_{\text{games}} = 144$)			
SV (Shapley, 1953)	30.09 ± 14.19	36.81 ± 48.40	0.629 ± 0.314
LOO (Covert et al., 2021)	26.41 ± 12.95	34.72 ± 47.77	0.627 ± 0.314
FT (Cemri et al., 2025)	60.16 ± 20.70	20.01 ± 37.16	0.090 ± 0.368
FA (Zhang et al., 2025b)	75.52 ± 18.53	11.59 ± 29.91	-0.087 ± 0.402
ASV, half-edge retention	19.75 ± 16.74	48.61 ± 50.16	0.736 ± 0.289
ASV, relabelled graph	34.03 ± 23.79	38.89 ± 48.92	0.569 ± 0.379
ASV (ours) (Frye et al., 2020)	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000
Independent controls ($n_{\text{games}} = 36$)			
SV (Shapley, 1953)	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000
LOO (Covert et al., 2021)	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000
FT (Cemri et al., 2025)	37.58 ± 20.34	43.84 ± 48.36	0.391 ± 0.424
FA (Zhang et al., 2025b)	48.41 ± 18.37	29.14 ± 42.62	0.220 ± 0.386
ASV, half-edge retention	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000
ASV, relabelled graph	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000
ASV (ours) (Frye et al., 2020)	0.00 ± 0.00	100.00 ± 0.00	1.000 ± 0.000

Metrics and aggregation. Using $N(s) = s/\|s\|_1$ with $N(0) = 0$, define $p_m = N(s_m)$ for exact methods and $p_m = [N(s_m^{(1)}) + N(s_m^{(2)})]/2$ for FT/FA. The normalized recovery error is

$$\text{Err}_m = \tfrac{1}{2}\|p_m - N(\boldsymbol{\delta})\|_1. \tag{40}$$

Top-1 accuracy assigns credit $1/k$ if the true maximum-contribution action is among k tied predicted maxima, and zero otherwise; numerical ties use absolute tolerance 10^{-12}. Spearman correlation uses average ranks, assigning zero to constant predictions. Error and Top-1 are reported as percentages. Means and sample standard deviations are computed across games; repeated model calls are not counted as independent games.

Table 13: Mean normalized recovery error (%) by graph family; 36 games per row.

Family	SV	LOO	FT	FA	ASV (ours)
Chain	36.40	27.93	55.70	83.52	0.00
Fork	41.52	37.88	61.80	75.56	0.00
Join	10.39	10.46	68.52	79.71	0.00
Random DAG	32.04	29.38	54.62	63.28	0.00
Independent	0.00	0.00	37.58	48.41	0.00

Results and verification. Tables 3, 12, and 13 show that ASV recovers the intervention-derived contributions to numerical precision throughout the dependent games, while SV and LOO have mean errors of 30.09% and 26.41%. All three exact methods recover the independent controls. ASV's error increases to 19.75% with incomplete graphs and 34.03% with relabelled graphs, supporting the role of correct causal structure in this recovery task. FT/FA have larger errors under this prompting protocol; their scores also reflect numerical lookup interpretation. An independent rational implementation checks all 540 primary exact vectors and explicitly enumerates legal orders for 180 graph-specific vectors in the six-action games. Independent Decimal aggregation verifies the reported metrics from raw model responses, and all 720 request hashes match the frozen prompts.