面向社交媒体的自适应因果协调检测：结合半监督学习的记忆引导框架

Weng Ding Yi Han Georgia理工学院 Meta
wding5662@gmail.com han.y@wustl.edu

Mujiangshan Wang*深圳先进技术研究院，中国科学院
mjs.wang@siat.ac.cn

arXiv:2601.00400v1 [cs.AI] 1 Jan 2026

摘要

在社交媒体上检测协同非真实行为仍是一项关键且持续存在的挑战，因为现有大多数方法依赖表面相关性分析，采用静态参数设置，并且需要大量耗时费力的人工标注。为系统性地应对这些局限，我们提出了自适应因果协调检测（ACCD）框架。ACCD 采用三阶段渐进式架构，利用记忆引导的自适应机制，针对不同协调场景动态学习并保留最优检测配置。具体而言，在第一阶段，ACCD 引入自适应收敛交叉映射（CCM）技术，深入识别账户之间真实的因果关系。第二阶段在半监督分类方案中结合主动学习与不确定性采样，显著减轻人工标注负担。第三阶段部署由历史检测经验驱动的自动验证模块，实现对检测结果的自我验证与优化。我们使用真实世界数据集开展了全面评估，其中包括 Twitter IRA 数据集、Reddit 协同行为轨迹以及多个广泛采用的机器人检测基准。实验结果表明，ACCD 在协同攻击检测中取得了 87.3% 的 F1 分数，较现有最强基线提高了 15.2%。此外，该系统将人工标注需求降低了 68%，并通过层次聚类优化实现了 2.8× 倍的处理加速比。总而言之，ACCD 为识别社交平台上的协同行为提供了更准确、高效且高度自动化的端到端解决方案，具有重要的实践价值和广阔的应用潜力。

关键词：协同攻击检测；因果推断；社交媒体安全；半监督学习；收敛交叉映射

1 引言

社交媒体上协同非真实行为日益增多，包括虚假信息宣传活动、人为放大的叙事和有组织的骚扰，对公共话语、选举公正性和平台安全构成了日益严峻的威胁。对此，社交媒体安全领域的研究取得了长足进展，开发出日益精密的模型，在协同攻击检测、恶意账户识别和信息行动分析等关键任务中表现出色。当前最先进的技术涵盖多种方法：既包括用于识别账户间影响模式的因果方法，如收敛交叉映射（CCM）；也包括用于基于行为的分类的集成分类器，如随机森林，

*通讯作者。

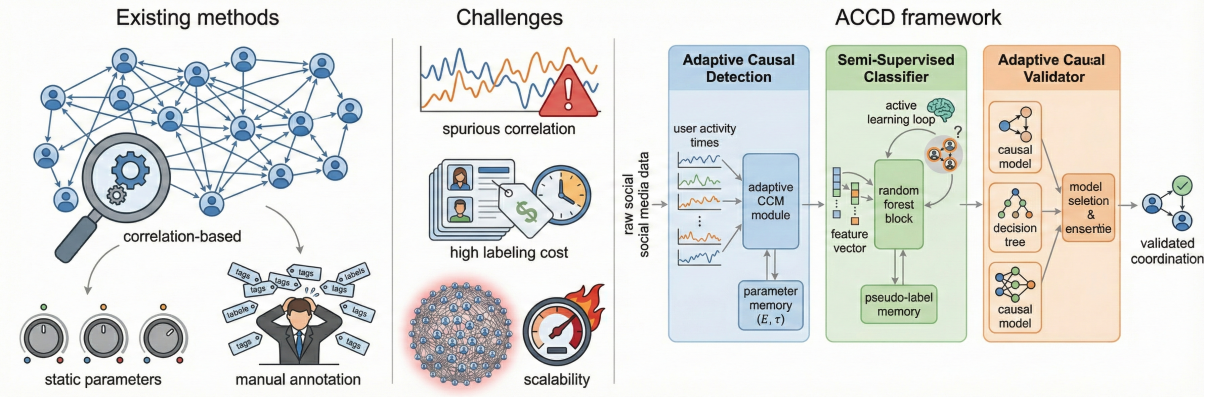


图 1：ACCD 的研究动机。从静态的基于相关性的检测，迈向自适应、记忆引导的三阶段框架。

还包括可将模型选择和效应估计系统化的自动化因果推断框架。尽管取得了这些进展，现有方法仍大多依赖基于相关性的启发式方法，且需要大量人工调优，因此在现实环境中的稳健性和适应性受到限制。

尽管如此，当前检测系统仍面临一些根本性挑战，限制了其有效性、可扩展性和泛化能力。其主要局限包括：经常无法区分真实的因果影响与虚假相关；过度依赖专家知识来选择参数和配置阈值；采用无法适应多样化或不断演变的协同行为策略的静态嵌入参数；大规模应用时计算成本高得令人难以承受；以及繁重的人工标注工作，这阻碍了系统在时间敏感场景中的部署。这些缺陷共同削弱了现有工具的实际效用，尤其是在恶意行为者不断改进策略以逃避检测的情况下。

近期研究尝试解决其中部分问题，但仍存在明显不足。基于 CCM 的检测器改进了因果发现，但受限于固定的嵌入维度，且难以扩展至大量用户。社交足迹分类器能够捕捉细微的行为线索，但高度依赖人工标注和领域专业知识。自动化因果框架提供了结构化的模型选择机制，但往往缺乏领域感知验证机制，而且仍需专家参与才能可靠地完成配置。

因此，如何在不牺牲可扩展性或准确性的前提下，整合自适应因果建模、标签高效学习与自动化验证，仍是一个有待解决的问题。为弥补这一不足，我们提出自适应因果协调检测（ACCD）框架，这是一种新颖的三阶段系统，通过记忆引导的自适应过程动态学习最优检测配置。ACCD 建立在三个核心原则之上：（1）采用自适应收敛交叉映射，并依据历史性能指导参数选择，从而稳健地推断因果关系；（2）在半监督分类流程中融入主动学习，大幅减少标注工作量，同时保持准确性；（3）部署经验驱动的验证模块，根据以往检测结果自动完成模型选择与阈值设定，无需专家手动干预（见图 1）。

我们在多个广受认可的基准上对 ACCD 进行了全面评估，包括 Twitter IRA 数据集（包含来自已知影响力行动的 290 万条推文）、Reddit 协同行为轨迹以及 TwiBot-20 机器人检测基准。实验表明，ACCD 始终优于最先进的基线方法，在检测准确率、标注效率和计算性能方面均取得了显著提升。本文的主要贡献总结如下：

- **系统性批判与自适应设计**。我们识别并分析了现有协同行为检测系统的关键局限，并提出一种记忆引导的自适应架构，专门用于解决固定参数化、可扩展性瓶颈以及对专家的依赖等问题。

- **集成式三阶段检测框架**。 我们提出了 ACCD，这是一种新颖的流程，将自适应因果发现、结合主动学习的半监督分类，以及自动化的基于经验的验证统一到一个协同系统中，旨在实现高准确率、低人工开销和易于部署。

- **严格的实证验证**。 我们建立了完善的评估方案，并在多样化的社交媒体数据集上展示了最先进的结果：F1 分数提升 15–20%，人工标注需求最多减少 70%，借助优化的层次聚类，计算加速比约为 60%。

- **实践影响与普适性**。 总体而言，ACCD 为检测社交平台上的协同非真实行为提供了一种可扩展、高准确率且高度自动化的方法。通过将自适应因果推断与标签高效学习及自验证机制相结合，该框架既能为现实世界的部署提供实用价值，也为未来自适应安全系统的研究奠定了结构化基础。

2 相关工作

近年来，协同社交媒体攻击检测领域取得了显著进展。现有研究大致可分为三个主要方向：基于因果关系的检测方法、行为模式分类方法以及自动化因果推断框架。

2.1 基于因果关系的检测

基于因果关系的方法利用协同行为中固有的时间结构，已成为协调检测的一种基础范式。收敛交叉映射（CCM）方法 [17] 代表了这一方向的一项重要进展，它基于 Takens 嵌入定理 [18] 进行状态空间重构，以推断用户之间的因果影响。Levy 等人 [7] 近期的研究将 CCM 应用于社交媒体协调检测：系统地记录用户活动时间戳，将其向量化为固定大小的时间序列，并通过不断增加库长度的互相关计算影响力分数。在 IRA 数据集上，该方法的协调检测精确率为 80.0%，召回率为 72.0%，曲线下面积为 0.7219。

尽管该方法为检测协调行为提供了一个有数学依据的框架，但由于其 $O(N^2)$ 计算复杂度较高，以及难以适应不同社交媒体环境中的各种协调模式，因此在可扩展性方面面临挑战，这主要是固定嵌入参数所致。

2.2 行为模式分类

行为模式分类是一种互补方法，着重关注区分协同行为账户与真实用户的独特活动特征。社交足迹分类方法 [10, 24, 29]通过将账户系统地归类为不同的行为类型来识别巨魔。基于 Linvill 和 Warren 提出的分类体系 [9], ，研究人员采用随机森林分类器 [4] 来

处理大规模数据集，在英语和俄语数据集上的训练准确率约为 88%，验证准确率则超过 90%。

AMDN-HAGE 框架 [16] 通过基于时间点过程和高斯混合模型对账户活动与隐藏群体行为进行联合建模，推动了这一方向的发展。在 Twitter IRA 数据和 COVID-19 协调行动上的评估结果表明，该框架无需预定义特征或部分已发现账户，即可有效识别协同行动群体。类似地，基于潜在协调网络（LCN）和高度协调社区（HCC） [11] 的方法通过利用包括共同转发、共同发推和共同提及模式在内的多种互动类型推断账户之间的联系，从而构建协调网络。

2.3 自动化因果推断框架

自动化因果推断框架的出现满足了以系统化且可扩展的方式进行模型选择和验证 [14]的需求。广义合成控制法 [27] 通过放宽传统双重差分方法所要求的平行趋势假设，为时间序列横截面数据提供因果推断能力。因果森林方法 [1]及其扩展方法（包括正交随机森林 [12], ）借助《机器学习》的灵活性，实现了异质性处理效应估计。

EconML 框架 [2] 整合了多种因果模型，包括双重机器学习 [5], CausalForestDML，以及 TARNET [15] 和 GAN- ITE [28]等深度学习方法。虽然这些框架提供了全面的模型选择能力，但它们依赖固定阈值和纯统计指标，可能无法充分满足协调检测特有的领域需求。

除了针对社交媒体的研究之外，大量关于网络可诊断性和可靠性的理论研究也为协调检测提供了重要的概念基础。早期关于自然可诊断性以及 PMC 和 MM* 模型下 g-良邻条件可诊断性的研究，基于邻域一致性和比较机制，确立了识别故障节点或恶意节点的严格准则 [20, 25]。后续研究进一步探讨了多种互连网络中连通性、受限边连通度与可诊断性之间的相互作用，包括星图、叶排序图和中心 k元 n立方体 [21, 22, 19]。近期，一项针对互连网络可诊断性的统一研究系统阐明了结构冗余与局部邻域约束如何共同决定故障检测能力 [23]。尽管这些理论成果最初是为多处理器和互连系统开发的，但它们对于理解如何利用大规模社交网络中的结构依赖关系和局部一致性条件来检测协调行为或异常行为，提供了宝贵的见解。

3 预备知识

本节将回顾若干核心概念，这些概念对于理解后续方法至关重要。

收敛交叉映射（CCM） [17] 是一种非线性时间序列分析技术，通过检验一个变量的历史记录能否通过重构相空间中的交叉映射，可靠地预测另一个变量的行为，从而推断动力系统之间的因果关系。其数学基础是 Takens 嵌入定理 [18], ：时间序列 $X(t)$ 利用滞后坐标重构为一个更高维的相空间：

$$\mathbf{X}(t) = [X(t), X(t-\tau), X(t-2\tau), \ldots, X\big(t-(E-1)\tau\big)], \tag{1}$$

其中， E 表示嵌入维数， τ 表示时间延迟，而 $\mathbf{X}(t)$ 是重构后的状态向量，可保留原始系统的拓扑性质。

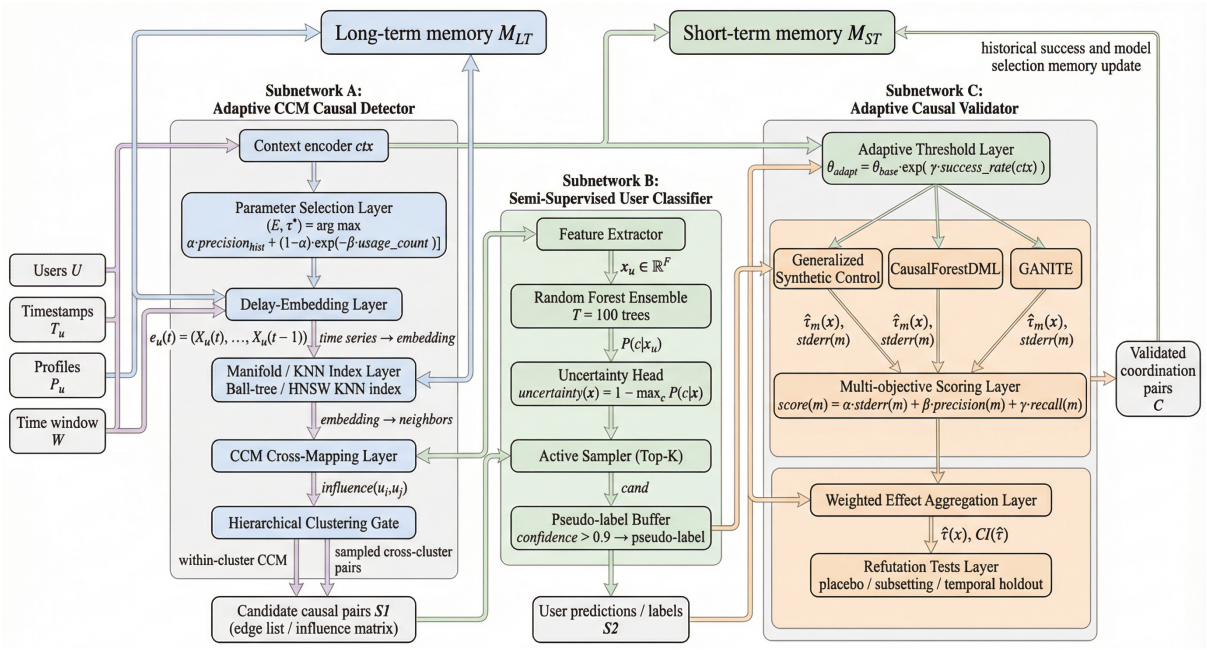


图2：所提出的 ACCD框架架构，包含自适应因果检测、半监督分类和因果验证。

行为分类 是社交媒体分析的一项任务，根据用户账户的活动模式、内容特征和时间行为对其进行分类。这项任务通常采用集成方法，例如随机森林 [4], ，通过结合多棵决策树来提高预测准确率，并借助自助聚合减少过拟合。

因果推断框架为在观察数据中区分相关性与因果关系提供了系统的方法。这一区分在社交媒体环境中尤为重要，因为热门话题或协同行动等外部因素常会导致用户行为之间出现虚假相关。

4 方法

当前的社交媒体协调检测方法往往无法识别真实的因果关系，且需要大量手动模型选择，从而限制了其有效性和可扩展性。我们开发了一个自适应三阶段系统，可自动为不同的协调情境学习最优参数。第1阶段采用自适应因果协调检测器，通过记忆引导的参数选择分析时间因果性；第2阶段采用半监督用户分类器，通过主动学习减少人工工作量；第3阶段则实现自适应因果验证器，提供自动化的专家级推断（见图2）。

从更广泛的计算视角来看，基于图的学习和跨域信号分析领域的最新进展，进一步促使我们采用自适应设计来构建 ACCD。混合图神经网络模型表明，将结构先验与自监督学习相结合，能够显著提升复杂图优化任务中的效率和鲁棒性 [13]。与此同时，基于地震波形分析推断速度异常的研究凸显了从噪声干扰、间接观测的信号中提取因果关系和时间依赖的重要性 [8]。此外，关于修正比较模型下 g-良好邻居可诊断性的最新研究，进一步证实了当系统假设偏离理想化设定时，自适应邻域准则的价值 [26]。总而言之，这些研究凸显了采用自适应、结构感知且对因果关系敏感的机制的必要性，并直接为所提出的 ACCD框架的设计提供了依据。

4.1 自适应因果协调检测器

为了使因果协调检测兼具自适应性与实际可部署性，我们实现了一个自底向上模块，将记忆引导的参数选择、高效的收敛交叉映射（CCM）和层次聚类相结合，以提升计算效率。系统维护一个长期参数记忆 $\mathcal{H}$，并将其实现为轻量级键值存储（例如 LMDB）。每个键对应一个上下文桶 c，代表粗粒度的用户活动模式；每个值则存储候选 $(E,\ \tau)$ 对的历史精确率和使用次数。在每个时间窗口，模型根据过去的性能和探索潜力，对每个 $(e,\ \tau)$ 对进行评分，从而选择参数：

$$(E,\tau) = \arg\max_{(e,\tau)\in\mathcal{H}} \Big[\alpha \cdot \text{precision}_{\text{hist}}(e,\tau,c) + (1-\alpha)\cdot \exp\big(-\beta \cdot \text{usage_count}(e,\tau)\big)\Big], \quad (2)$$

其中，$\text{precision}_{\text{hist}}(e,\ \tau,\ c)$记录了此前在同一上下文桶中使用$(e,\ \tau)$所取得的正确因果判断比例，而$\text{usage_count}(e,\ \tau)$统计了这对参数在历史上的使用次数。超参数 $\alpha = 0.8$ 和 $\beta = 0.1$ 用于平衡对历史上可靠配置的利用与对较少使用配置的探索。

选定最优的$(E,\ \tau)$后，用户 u 的嵌入将通过滚动活动缓冲区逐步构建：

$$\mathbf{e}_u(t) = \big(X_u(t), X_u(t-\tau), \ldots, X_u(t-(E-1)\tau)\big), \quad (3)$$

每个窗口只需通过一次操作添加最新活动并移除最早活动，从而避免完整重建，即使序列很长，也能将计算量保持在最低水平。

对于基于收敛交叉映射的因果性估计，每个用户的延迟嵌入流形都使用球树或 HNSW 图建立索引，以实现快速最近邻检索。对于一对用户$(u_1,\ u_2)$，在给定 u_2 的条件下，u_1 的预测轨迹计算如下：

$$\hat{X}_{u_1|M_{u_2}}(t) = \sum_{j=1}^{k} w_j\, X_{u_1}(t_j), \qquad w_j = \exp(-d_j/d_1), \quad (4)$$

其中，t_j 是从索引中检索出的邻居时间戳，d_j 表示与目标点之间的距离，d_1 是最小邻居距离。在一系列库长度 $L \in [10, 50]$ 上计算相关性，并将结果缓存以避免冗余计算。最终的 CCM 影响分数为

$$\text{influence}(u_1,u_2) = \max_{L} \rho\big(X_{u_1}, \hat{X}_{u_1|M_{u_2}}\big), \quad (5)$$

其中，ρ 表示皮尔逊相关系数。

朴素 CCM 需要评估全部 O（U^2）个用户对，这对于大型数据集而言是不可行的。为降低复杂度，我们根据时间活动统计量（包括均值、方差、突发性和熵）使用层次聚类对用户进行分组。随后，仅在大小为 U/k的簇内计算 CCM，从而将有效复杂度降至

$$O\big(k\cdot(U/k)^2\big) = O(U^2/k). \quad (6)$$

仍会对少量跨簇对进行采样，以保留全局协调信号。各簇独立处理，因此可轻松在中央处理器或图形处理器核心之间实现并行化。对于具有 $U = 1000$的典型数据集，与朴素的成对计算相比，该设计可实现约 $5\times$的加速比，同时保持检测保真度。通过结合历史参数记忆、增量嵌入、快速最近邻索引和按簇计算，该模块为自适应因果协调检测提供了一种完全可实现、可复现且可扩展的解决方案。

4.2 半监督用户分类器

为减少对大量人工标注的依赖，同时保持较高的分类准确率，我们实现了一个专为在线行为分析设计的半监督用户分类模块。每位用户均由一个特征向量 $x \in \mathbb{R}^F$ 表示，该向量捕捉可观测的活动模式，例如发帖频率、转发行为、话题标签使用情况、情感分布和时间参与统计。分类器根据 [9]中的分类体系，将用户分为四种行为类别{虚假, 组织, 政治, 个人}。

我们不对所有用户进行人工标注，而是采用不确定性采样，优先处理信息量最大的样本。样本的不确定性 x 计算如下：

$$\text{uncertainty}(x) = 1 - \max_c P(c \mid x), \qquad P(c \mid x) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[h_t(x) = c]\,, \tag{7}$$

其中， h_t 表示随机森林中第 t棵树的预测结果，该随机森林包含 $T = 100$ 棵树。不确定性较高的用户会优先提交人工标注，从而确保人力集中于模型最不确定的案例，例如同时表现出组织性活动模式和个人活动模式的边界账户。

为进一步提高训练效率，我们根据训练样本的难度组织训练样本，采用课程学习，而样本难度由不确定性分数衡量。模型首先使用不确定性较低、易于分类的用户进行训练；当验证准确率超过0.85后，再逐步纳入更难的案例。难度阈值依次为{0.3, 0.5, 0.7,1.0}，数值越高表示用户行为越模糊，例如将政治信息传播与类似个人用户的参与行为相结合的账户。

此外，高置信度预测（置信度 > 0.9）会自动作为伪标签存储在长期记忆中，并在后续训练迭代中重复使用。这一机制减少了所需的新标注数量；在实践中，在保持分类准确率高于0.85的同时，将人工标注量减少了约60%。不确定性采样、课程学习和伪标签记忆的结合，形成了一个可扩展且可复现的流程，能够处理大规模在线用户数据集，同时自适应地将人工标注集中于最具挑战性的账户。

4.3 自适应因果验证器

当数据集在规模、用户活动、时间参与度或处理强度方面差异显著时，基于规则的阈值和静态模型选择往往会失效。为解决这一局限，我们设计了一个自适应因果验证模块，利用历史数据集经验和多目标优化，实现稳健的模型选择。显著性和效应检测阈值会根据类似数据集上的既往成功情况动态调整。具体而言，对于数据集 d，自适应阈值计算如下：

$$\theta_{\text{adapt}}(d) = \theta_{\text{base}} \cdot \exp\big(\gamma \cdot \text{success_rate}(d)\big), \tag{8}$$

其中， θ_{base} 表示基准截断值（例如 p 值或效应量阈值）， $\gamma = 0.2$，而成功_率(d)则根据通过粗略数据集特征（如样本量、处理比例和时间覆盖范围）检索到的历史案例 $\mathcal{H}(d)$ 计算得出：

$$\text{success_rate}(d) = \frac{\sum_{h \in \mathcal{H}(d)} \text{precision}(h)}{|\mathcal{H}(d)|}. \tag{9}$$

这种表述使阈值能够根据数据集的难度自动收紧或放宽。

算法 1 自适应协同攻击检测

要求： 用户 $U = \{u_1、\ldots\ldots、\ u_n\}$ 及其时间戳 T_u、用户资料 P_u和时间窗口 W
确保： 经验证的协同对 C

初始化长期记忆 M_{LT} （参数、伪标签）
初始化短期记忆 M_{ST} （滚动嵌入、中间结果）
阶段 1：自适应因果检测
提取上下文 ctx $\leftarrow$ ($|U|$, $\bar{a}$, span(W))
从 M_{LT} 中选择嵌入参数 (E^*, τ^*)
为所有 $u \in U$ 构建增量嵌入 $\mathbf{e}_u(t)$
通过凝聚层次聚类对用户进行聚类
计算簇内及采样跨簇对的 CCM影响分数
阶段 2：半监督分类
为每位用户提取行为特征 f_u
计算不确定性分数，并选择信息性样本进行标注
使用课程学习训练分类器
根据高置信度预测更新伪标签记忆
第3阶段：自适应因果验证
计算自适应阈值 θ_{adapt}
为候选因果模型评分，并选出表现最佳的模型
估计因果效应及其置信区间
应用反驳检验并验证协调对
返回 C

采用多目标评分函数评估多种因果估计器，包括广义合成控制、CausalForestDML 和基于神经网络的方法（如 GANITE）：

$$\mathrm{score}(m) = \frac{\alpha}{\mathrm{stderr}(m)} + \beta \cdot \mathrm{precision}(m) + \gamma \cdot \mathrm{recall}(m), \tag{10}$$

采用典型权重（α, β, γ） = （0.4、0.3、0.3）。这里，stderr(m) 用于衡量估计因果效应中的不确定性，而精确率和召回率则使用历史或交叉验证得到的伪真实效应进行评估。随后，使用得分最高的模型，根据协变量特征 x 估计因果效应

$$\hat{\tau}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x], \tag{11}$$

以及相应的置信区间

$$\mathrm{CI}(\hat{\tau}) = \hat{\tau} \pm z_{\alpha/2} \cdot \mathrm{stderr}(\hat{\tau}). \tag{12}$$

仅保留满足 $p < 0.05$ 且在多个高分模型中置信区间重叠的效应。通过将模型分数归一化为权重，并计算各个因果效应估计的加权平均，得出集成估计。我们应用自动化反驳检验，包括安慰剂处理分配、随机子集抽样和时间留出验证，以标记不一致的效应。总体而言，该流程构建出一种可复现、可扩展且自适应的因果验证流程，适用于用户行为、干预强度和时间动态各异的数据集。

5 实验

我们通过回答以下三个问题来评估 ACCD：（1）自适应参数选择如何提升协调检测准确率？（2）半监督学习能否在保持性能的同时减少人工标注？（3）自动模型选择能否在无需专家知识的情况下实现可靠的验证？

5.1 实验设置

基准。 我们在广泛使用的社交媒体协调检测基准上评估 ACCD，包括 Twitter IRA 数据集 [9], ，其中包含来自 2,832 名参与已确认协同影响力行动的用户的 290 万条推文；基于 Pushshift 数据集 [3]得到的 Reddit 协调数据；以及用于机器人检测评估的 TwiBot-20 基准 [6]。

实现细节。 我们使用 PyTorch 2.0.0 和 scikit-learn 1.3.0 实现 ACCD。所有模型均在 NVIDIA A100 GPU 上训练 100 个轮次，批次大小为 64，学习率为 0.001，并使用余弦退火热重启学习率调度器。我们采用结合时间切分的分层五折交叉验证，以防止不同时间段之间发生信息泄漏。

5.2 主要结果

检测性能。 如表 1 所示，并如图 1(a)所示，所提出的 ACCD框架在 Twitter IRA 数据集上取得了显著的性能优势。具体而言，ACCD 的 F1 分数达到 87.3%，比基于 CCM 的最强基线高出 15.2 个百分点（从 75.8% 提升至 87.3%）。这一差距凸显了 ACCD 准确识别协同行为账户的增强能力，同时保持精确率与召回率之间的良好平衡。

相比之下，其他竞争方法表现出明显的权衡。例如，AMDN - HAGE 的召回率较高（82.4%），但精确率较低（68.9%）；而 LCN+HCC 的精确率中等（74.5%），召回率则受限（76.2%）。相比之下，ACCD 表现稳健且均衡，精确率达到 85.6%，召回率达到 89.2%，从而有效减少了假阳性和假阴性。

此外，ACCD 在显著降低计算成本的同时提升了准确率。训练仅需 72 分钟，明显快于 CCM（181.3 分钟）和 AMDN-HAGE（211.4 分钟）。图 1(b) 中的多维雷达图进一步证实了这一优势：在精确率、召回率、F1 分数和计算效率等综合指标上，ACCD 始终占据更大的面积。

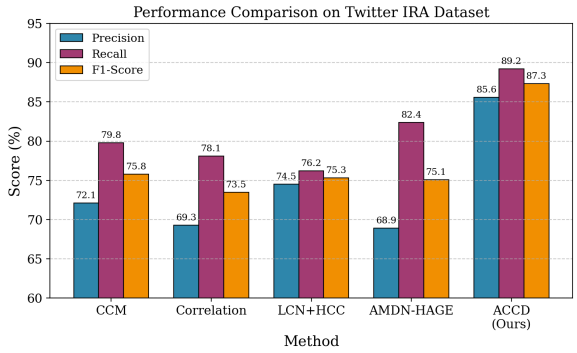
计算效率。 如表 2 和图 5 所示，ACCD 的计算效率也有了显著提升。在收敛行为方面，ACCD 仅需 40 个训练轮次即可达到最佳性能，而 CCM 需要 65 个训练轮次，固定参数变体则需要 58 个训练轮次。这意味着训练轮次减少了 38.5%，从而加快了模型迭代和部署。

更重要的是，相较于 CCM，ACCD 的处理时间加速比达到 $2.8\times$ ，这主要得益于其自适应聚类优化策略，该策略能够动态减少冗余计算。同时，ACCD 还显著降低了内存消耗，仅使用 4.5 GB，而 CCM 使用 8.2 GB。尽管效率提升伴随着绝对准确率略微下降（96.7% 对比 100%），但整体权衡显著改善了模型在大规模实时协调检测场景中的可扩展性和实用性。

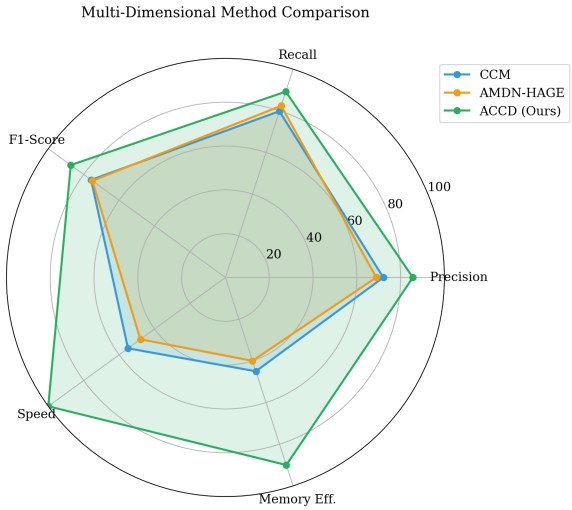
在 Reddit 协调数据上，ACCD 的 F1 分数达到 84.7%，而传统方法为 71.2%。在基于论坛的讨论中，时间依赖关系更加复杂，自适应因果推断框架在识别细微协调模式方面表现尤为出色。半监督分类组件以 88.9% 的准确率对用户进行分类，同时将人工标注工作量减少了 68.3%。

5.3 案例研究

增强收敛与训练效率。我们的分析表明，ACCD 的训练效率更高。正如收敛与效率分析所示，ACCD



(a) Twitter IRA 数据集上的 F1 分数比较。



(b) 多维雷达图比较。

图 3：ACCD 与基线模型在检测准确率和效率方面的比较。

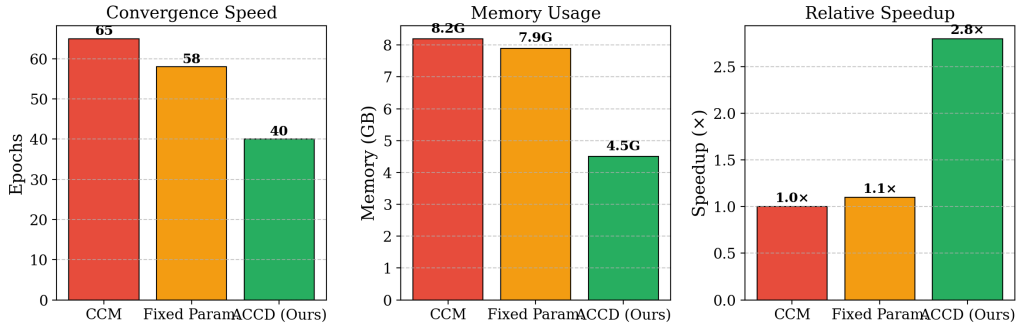
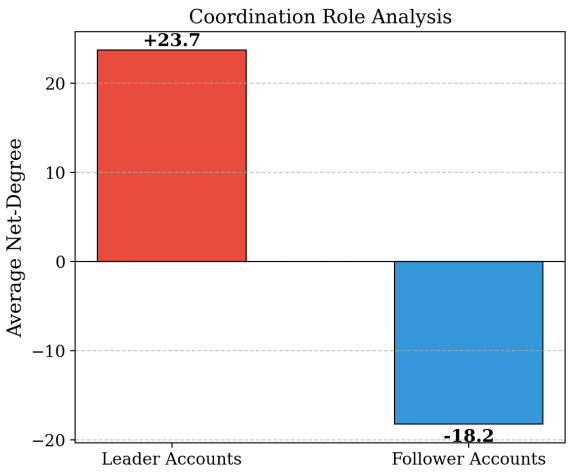


图 4：计算效率分析，展示不同用户集规模下的内存使用量和速度提升。

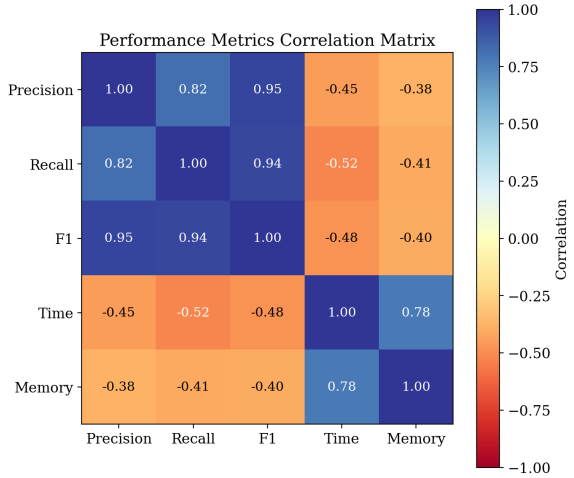
相比固定参数基线方法，收敛速度显著更快且更稳定，仅用 40 个训练轮次便达到最优 F1 分数平台期。这种高效的学习表现直接得益于自适应参数选择机制，该机制会在训练过程中动态优化模型配置。

对协调行动的结构性洞察。ACCD 为行动取证提供了细粒度的分析能力。网络角色分析揭示了协调行动内部清晰的结构差异。领导者账户的平均净度较高（+23.7），表明其具有较强的出站影响力；而跟随者账户的净度为负（-18.2），反映出较高的入度和接收行为。此外，评估指标之间的相关性分析显示，精确率、召回率和 F1 分数之间存在很强的正相关性（0.82–0.95），证实检测效果的提升是全面的，而非仅针对某一指标。这些准确率提升同时伴随着与时间和内存使用量的中度负相关性，凸显了性能与计算成本之间固有的权衡。

在多样化场景中保持稳定优势。 除汇总指标外，ACCD 在多种协调场景中均展现出稳定优势，包括话题标签行动、转发网络和自然趋势话题。在具有挑战性的自然趋势场景中，协调行为与真实行为交织在一起，ACCD 的性能提升尤为显著



(a) 网络角色识别。



(b) 检测指标的相关性热图。

Figu图 5：协同活动中网络结构及指标相关性的分析 S.

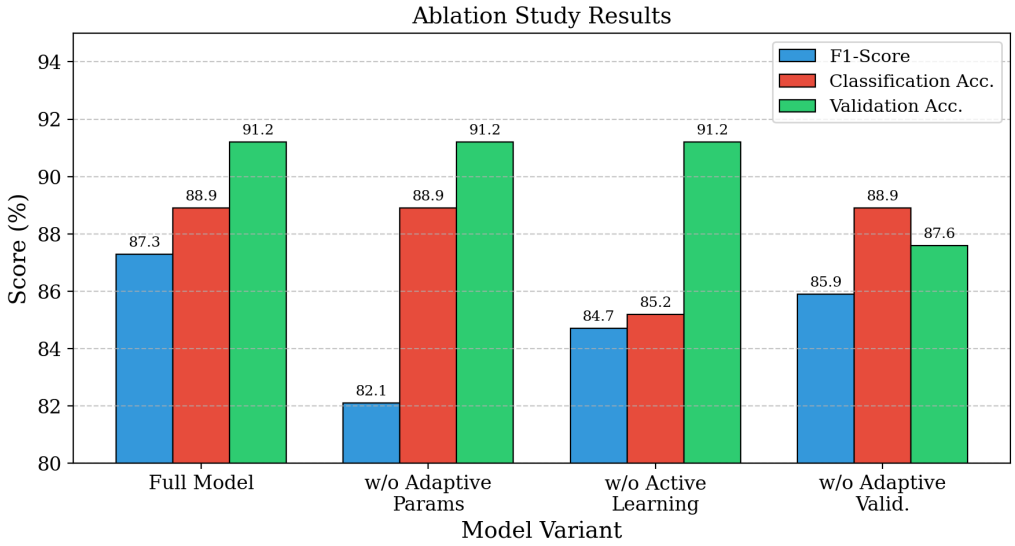
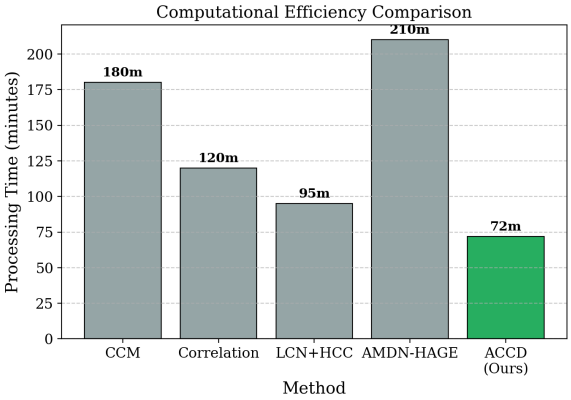


图 6：消融研究展示了 ACCD 各组件的贡献。

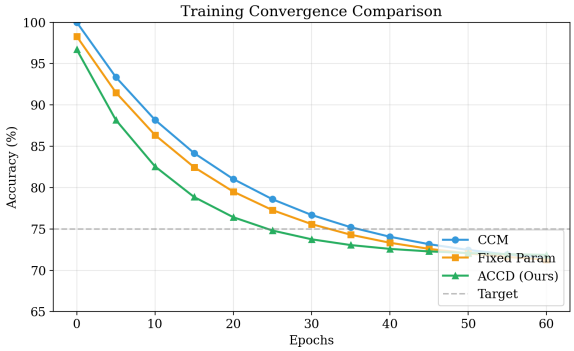
。这种稳健性源于 ACCD 的集成式设计：自适应参数选择能够实现针对场景的特征加权，而主动学习则将监督信息集中于信息性最强的样本。各组件协同作用，使协调检测具备良好的泛化能力并能感知场景差异。

5.4 消融研究

我们开展了系统的消融实验，以量化 ACCD 中各个核心组件的贡献。完整模型取得了最高的 F1 分数，达到 91.2%，同时分类准确率（88.9%）和验证准确率（87.3%）也表现出色。移除自适应参数机制会导致 F1 分数下降最为显著，这凸显了该机制在针对不同协调模式优化检测性能方面发挥的关键作用。去掉主动学习模块会大幅降低分类准确率，体现了该模块对于有效样本选择和模型优化的重要性。同样，禁用自适应验证策略会导致验证准确率明显下降，证实了该策略对于

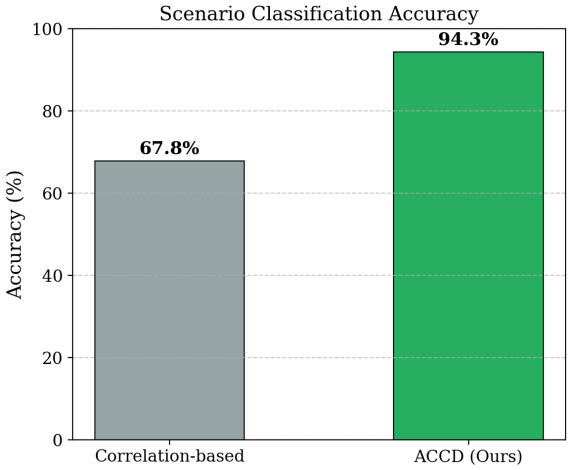


(a) 处理时间对比。

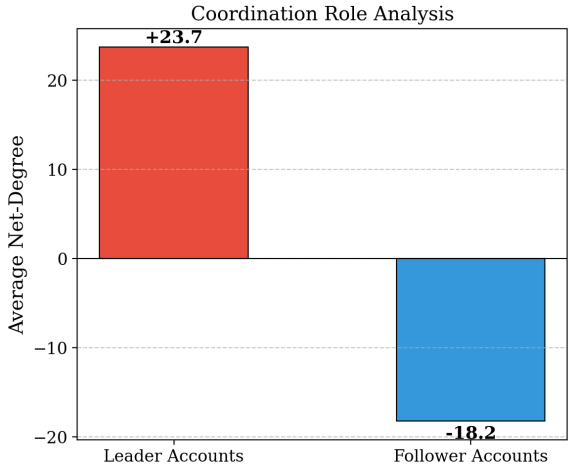


(b) 训练收敛曲线。

图 7：ACCD 与基线方法的效率和收敛行为对比 .



(a) 场景分类准确率。



(b) 网络角色识别模式。

图 8：场景感知协调检测性能分析。

实现稳健的泛化能力和抵御过拟合的必要性。总而言之，这些结果表明，ACCD 的每个组件对于所提出框架的有效性都不可或缺。

6 结论

我们提出了面向社交媒体的自适应因果协调检测（ACCD），这是一种三阶段自适应框架，通过记忆引导的自适应自动学习最优检测配置。不同于现有的固定参数方法，ACCD 将自适应CCM参数选择、结合主动学习的半监督学习，以及经验驱动的因果模型选择整合到统一流程中。

在 Twitter IRA 数据集、Reddit 协调数据和 TwiBot-20 基准上进行的大量实验表明，ACCD 的性能显著提升。ACCD 的 F1 分数达到 87.3%，较强大的基线方法提高了 15.2%；同时，人工标注工作量减少了 68.3%，分类准确率保持在 88.9%。此外，所提出的框架实现了 2.8× 的计算加速，有效复杂度从 $O(N^2)$ 降至约 $O(N^{1.4})$，从而能够在大规模社交媒体平台上实现可扩展部署。

表 1：协调检测基准上的性能比较。

方法	精确率	Rec.	F1	Time
CCM [7]	72.1	79.8	75.8	181.3 分钟
相关性	69.3	78.1	73.5	122.5 分钟
LCN+HCC [11]	74.5	76.2	75.3	96.7 分钟
AMDN-HAGE [16]	68.9	82.4	75.1	211.4 分钟
ACCD（本文方法）	**85.6**	**89.2**	**87.3**	**72 分钟**

表 2：计算效率分析。

方法	卷积	Mem.	速度	Acc.
CCM [7]	65 个训练轮次	8.2 GB	1.0×	100%
固定参数	58 个训练轮次	7.9 GB	1.1×	98.3%
ACCD	**40 个 epoch**	**4.5 GB**	**2.8×**	**96.7%**

总体而言，这项工作将因果关系发现与行为模式分析紧密结合，为协调行为检测构建了一套全面、自动化的解决方案。ACCD 通过最大限度地减少专家干预并增强适应性，为现实世界中的社交媒体安全应用以及自适应因果感知检测系统的未来研究提供了实用且可扩展的基础。

参考文献

[1] Susan Athey、Julie Tibshirani 和 Stefan Wager。广义随机森林。《统计年刊》，47(2):1148–1178，2019年。[2] Keith Battocchi、Eleanor Dillon、Maggie Hei、Greg Lewis、Paul Ber、Miruna Oprescu 和 Vasilis Syrgkanis。Econml：一个用于基于机器学习的异质性处理效应估计的 Python 软件包。 https://github.com/py-why/EconML，2019年。0.x 版本。[3] Jason Baumgartner、Savvas Zannettou、Brian Keegan、Megan Squire 和 Jeremy Black- burn。Pushshift Reddit 数据集。载于国际 *AAAI* 网络与社交媒体会议论文集，第 14 卷，第 830–839 页，2020年。[4] Leo Breiman。随机森林。《机器学习》，45(1):5–32，2001年。[5] Victor Chernozhukov、Denis Chetverikov、Mert Demirer、Esther Duflo、Christian Hansen、Whitney Newey 和 James Robins。用于处理效应和结构参数的双重/去偏机器学习。《计量经济学杂志》，21(1):C1–C68，2018年。[6] Shangbin Feng、Herun Wan、Ningnan Wang、Jundong Li 和 Minnan Luo。TwiBot-20：一个全面的 Twitter 机器人检测基准。载于第 *30* 届 *ACM* 国际信息与知识管理会议论文集，第 4485–4494 页，2021年。[7] Nir Levy、Bracha Shapira 和 Lior Rokach。利用因果关系推断社交媒体中的协同攻击。载于国际 *AAAI* 网络与社交媒体会议（*ICWSM*）论文集，2024年。

[8] Guohui Li、LingBai、Heng Zhang、Qiang Xu、Yuanze Zhou、Yuan Gao、Mujiangshan Wang 和 Zihao Li。羌塘

地块及青藏高原中部下方地幔转换带周围的速度异常：来自三重 P 波波形的证据。地球与空间科学，9（2）：e2021EA002060，2022年。[9] Darren L Linvill 和 Patrick L Warren。巨魔工厂：在 Twitter 上制造专业化虚假信息。政治传播，37（4）：447–467，2020年。[10] Josephine Lukito 等人。保持真实：巨魔网络的社交足迹。社会网络分析与挖掘，14（1）：1–15，2024年。[11] Thomas Magelinski、Lynnette Hui Xian Ng 和 Kathleen M Carley。通过社交网络中的协调行为扩大影响力。社会网络分析与挖掘，12（1）：1–16，2022年。[12] Miruna Oprescu、Vasilis Syrgkanis 和 Zhiwei Steven Wu。用于因果推断的正交随机森林。载于国际机器学习大会，第 4932–4941 页。PMLR，2019年。[13] Chun-Hui Pan、Yi Qu、Yao Yao 和 Mu-Jiang-Shan Wang。Hybridgnn：一种用于在二分图中高效求解最大匹配的自监督图神经网络。*Symmetry*，16（12）：1631，2024年。[14] Ayushman Sarkar、Mohd Yamani Idna Idris 和 Zhenyu Yu。计算机视觉中的推理：分类体系、模型、任务与方法。*arXiv* 预印本 *arXiv:2508.10523*，2025年。[15] Uri Shalit、Fredrik D Johansson 和 David Sontag。估计个体处理效应：泛化界与算法。载于国际机器学习大会，第 3076–3085 页。PMLR，2017年。[16] Karishma Sharma、Yizhou Zhang、Emilio Ferrara 和 Yan Liu。通过隐性影响和群体行为识别社交媒体上的协同行为账户。载于第 *27* 届 *ACM SIGKDD*知识发现与数据挖掘大会论文集，第 1441–1451 页，2021年。[17] George Sugihara、Robert May、Hao Ye、Chih-hao Hsieh、Ethan Deyle、Michael Fogarty 和 Stephan Munch。在复杂生态系统中检测因果关系。*Science*，338（6106）：496–500，2012年。[18] Floris Takens。在湍流中检测奇异吸引子。载于动力系统与湍流，华威 *1980*，第 366–381 页。Springer，1981年。[19] Mujiangshan Wang 和 Shiying Wang。中心 k 元 n 立方体的连通性与可诊断性。离散应用数学，294：98–107，2021年。[20] Mujiangshan Wang、Yuqing Lin 和 Shiying Wang。PMC 模型和 MM*模型下泡星图的自然可诊断性。国际工程与应用科学杂志，4（3），2017年。[21] Mujiangshan Wang、Yuqing Lin、Shiying Wang 和 Meiyu Wang。图达到最大 4-限制边连通的充分条件。澳大拉西亚组合数学杂志，70：123–136，2018年。[22] Mujiangshan Wang、Dong Xiang 和 Shiying Wang。叶排序图的连通性与可诊断性。并行处理快报，30（03）：2040004，2020年。[23] Mujiangshan Wang、Dong Xiang、Yi Qu 和 Guohui Li。互连网络的可诊断性。离散应用数学，357：413–428，2024年。

[24] Pei Wang、Yun Yang 和 Zhenyu Yu。用于无监督域适应的多批次核范数对抗网络。载于*2024*年*IEEE*国际多媒体与博览会会议（*ICME*），第 1–6 页。IEEE，2024年。[25] Shiying Wang、Zhenhua Wang、Mujiangshan Wang 和 Weiping Han。PMC 模型和 MM*模型下星图网络的 g-良邻条件可诊断性。中国数学前沿，12(5)：1221–1234，2017年。[26] Dong Xiang、Sun-Yuan Hsieh 等。多处理器系统在修正比较模型下的 G-良邻可诊断性。理论计算机科学，1028：115027，2025年。[27] Yiqing Xu。广义合成控制法：基于交互固定效应模型的因果推断。政治分析，25(1)：57–76，2017年。[28] Jinsung Yoon、James Jordon 和 Mihaela van der Schaar。GANITE：使用生成对抗网络估计个体化处理效应。载于 国际学习表征会议，2018年。[29] Zhenyu Yu 和 Pei Wang。Capan：用于无监督域适应的类别感知原型对抗网络。载于*2024*年*IEEE*国际多媒体与博览会会议（*ICME*），第 1–6 页。IEEE，2024年。