arXiv:2603.11726v1 [cs.CE] 12 Mar 2026

部分可观测性下基于半老虎机反馈的针对生态对手的战略防御在线学习

Anjali Purathekandy [1] **Deepak N. Subramani** [1]

摘要

我们提出一种在线学习算法，用于针对行为模型未知且具有部分可观测性的战略性生态对手，计算自适应资源分配策略。我们的设定解决了安全博弈中的一个根本局限：当无法先验地对手的行为建模时，经典的基于均衡的方法便会失效。我们将该问题表述为组合动作空间中的遗憾最小化问题，并采用半老虎机反馈，其中收益在不同目标之间呈非平稳且相互依赖。我们的算法名为 HERDS（通过资源部署实现战略守卫的人象冲突缓解），在扰动领导者跟随（FPL）的基础上进行了三项创新：（1）通过观测到的损失驱动动态预算划分，实现探索与利用同步进行；（2）在攻击入口点无法识别的混杂观测下，自适应地估计收益；（3）采用与模型无关的学习方法，无需行为假设即可提供遗憾保证。我们在人象冲突缓解领域展示了这一框架；在该领域，具有智能的生态对手会表现出战略行为（最优觅食、空间记忆、自适应规避），但其行为模型难以处理。使用基于大象移动数据校准的基于智能体的模型进行的实验表明，与采用均匀探索的扰动领导者跟随（FPL-UE）相比，遗憾降低了 15–45%；面对自适应对手时，作物损害减少了 40–50%；且在 40–50 轮次内收敛，而基线算法需要 60–80 轮次。

1. 引言

人象冲突（HEC）仅在印度每年就造成约400人死亡和100头大象死亡，并通过作物侵袭直接影响50万个家庭（Fernando 等人，2008年；Shaffer 等人，2019年）。尽管社区采用了各种威慑手段（噪音、火、屏障）和主动守护，大象的高智商、空间记忆和适应能力使大多数缓解策略随时间推移而失效（Fernando 等人，2008年）。实施有组织的作物保护成本高昂，需要训练有素的人员、设备和经常性预算。

决策分析和博弈论已被证明在生物多样性管理和保护中具有重要价值（Frank&Sarkar，2010年）。实地经验表明，大象作为策略性对手发挥作用，但其行为模型未知。它们表现出最优觅食（以高价值作物为目标（Shaffer 等人，2019年））、风险最小化（夜间袭击（Fernando 等人，2008年））、空间记忆，以及利用安全缺口的适应性学习。然而，与可以从数据中推断对手行为的标准安全博弈不同，人象冲突呈现出一个根本悖论：我们在实地观察到明确的策略性行为证据，却缺乏任何关于大象作物侵袭选择的行为模型，也没有足够的大象与守卫互动数据来构建此类模型。不同地点之间显著的行为差异以及迅速变化的条件，进一步削弱了传统方法的可行性。

这个问题需要采用在线学习方法。战略性对手会调整战术，因此固定策略难以奏效，因为大象会适应可预测的巡逻路线。贝叶斯方法和基于模型的方法依赖于先验分布；鉴于缺乏行为模型且异质性很强，我们无法对这些分布做出可信的设定。相比之下，对抗性在线学习无需依赖准确的行为假设，就能针对任意对手策略提供遗憾保证。它还可以利用历史数据进行初始化，同时根据观察到的作物损害，自适应地引导探索工作。为了开发并评估结合安全博弈的在线学习策略，基于主体的模型（ABM）可提供安全的虚拟环境，用于测试策略，尤其是在实地实验风险较高时（McLane 等人，2011年；Kirkland 等人，2020年；Janssen 等人，2020年）。

[1]印度科学理工学院计算与数据科学系，班加罗尔，印度。通讯作者：Deepak N. Subramani <deepakns@iisc.ac.in>。

2026年作者版权所有

实地证据表明，守护森林—农业边界是最有效的缓解措施之一（Fernando 等人，2008年）。然而，目前的守护工作依赖彼此缺乏协调的个人行动，导致资源分配效率低下，部署重复且存在严重的安全缺口，聪明的大象会利用这些缺口。这带来了一个组合优化挑战：如何战略性地分配 K 名守卫到 N 个边界区段，以最大限度地减少作物损失，同时应对会学习并根据部署模式做出反应的自适应对手。

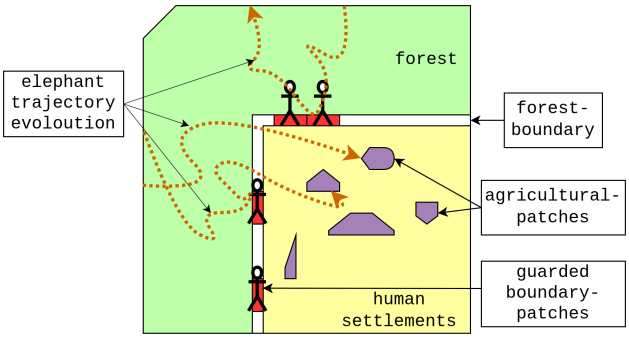

图1. 森林-农业边界沿线战略部署守卫以最大化拦截效果的示意图。图中展示了三种大象轨迹场景：（1）大象通过未设守卫的边界区段成功进入聚居地，造成作物侵袭；（2）大象在设有守卫的区段被有效拦截，折返森林；（3）尽管有护林员在场，大象仍通过另一处未设守卫的区段进入聚居地。黄色区域表示人类聚居地和农田；白色区域表示森林边界；绿色区域表示森林；红色符号标示守卫位置；棕色线条表示大象可能的路径。

我们将人象冲突缓解建模为绿色安全博弈，其中大象被视为战略对手，而防御方沿森林-农业边界分配有限的巡逻资源（图1）。这是首次对人象冲突进行博弈论分析，将安全博弈拓展至具有适应性战略行为的生态对手。由于缺乏可靠的行为模型，因此需要采用在线学习方法。为此，我们提出 HERDS（通过资源部署实现战略守卫的人象冲突缓解），该方法基于 FPL-UE 框架（Xu 等人，2016年），并包含三项创新：（1）根据观测到的作物损害动态调整探索与利用预算；（2）在部分可观测条件下自适应学习收益，同时考虑相互依赖的边界区段；（3）联合部署探索型和利用型巡逻。HERDS与模型无关，无需假设特定行为模型即可推断有效策略，并通过遗憾最小化提供性能保证。

我们的贡献包括：*(1)* 首个用于人象冲突的博弈论框架，在绿色安全博弈（GSG）结构中将大象建模为具有策略性的对手，并将安全博弈论扩展到具有未知行为模型的生态领域。*(2)*动态收益学习机制，即使受到部分可观测性约束，也能实时更新估计值。*(3)* HERDS算法将在线学习方法应用于存在未知对手的策略博弈。*(4)* 使用基于印度大象数据校准的基于主体的模型进行验证（Purathekandy 等人，2024年），证明了该方法相较于多种对抗模型具有更优的性能。

2. 相关工作

人象冲突缓解将安全博弈论（对策略性互动进行建模）与在线学习（处理未知的对手模型）相结合。相关工作涵盖三个领域：

在线学习基础。 在线学习处理在不具备完整环境知识情况下的序贯决策（Kalai & Vempala, 2005年; Cesa-Bianchi & Lugosi, 2006年）。基础性工作将遗憾最小化确立为对抗性环境中的核心性能准则。多臂老虎机算法为探索-利用权衡提供了次线性遗憾界（Auer 等人,2002年），并扩展到连续动作空间（Awerbuch & Kleinberg, 2008年）。部分监控框架处理不完整反馈（Audibert &Bubeck,2010年），而在线几何优化则应对自适应对手（McMahan & Blum,2004年）。

安全博弈中的在线学习。 当对手行为模型未知时，在线学习可在无需先验假设的情况下提供自适应解决方案。结合几何重采样的FPL算法处理半老虎机反馈下的在线组合优化（Neu&Bart´ok, 2013年）。FPL-UE通过均匀探索对此进行扩展，在无需收益或攻击者行为的先验知识的情况下实现可证明的次线性遗憾（Xu 等人, 2016年）。关键在于，无论对手策略如何，遗憾界均成立，这使其适用于无法先验可靠建模大象行为的人象冲突。近期工作在遗憾之外引入了重新分配时间标准（Zhu 等人, 2023年）以及用于计算多轮计划的方法（Marecki 等人, 2012）。

绿色安全博弈：框架基础。 我们的工作将绿色安全博弈扩展到野生动物保护。绿色安全博弈对防御方分配有限资源、以对抗瞄准高价值资产的对手进行建模（Fang 等人，2015年; Fang & Nguyen，2016年）。早期算法（ASAP（Paruchuri 等人，2007年））以及大规模应用（GUARDS、PROTECT（An 等人，2011年; Tambe，2011年））展现了实用价值。野生动物方面的应用包括采用偷猎者行为模型的PAWS（Yang 等人，2014年），以及使用POMDP表述的GMOP（Qian 等人，2014年）。然而，传统绿色安全博弈假设已知对手模型以计算均衡。人象冲突呈现出一个关键区别：大象是策略性的

对手（最优觅食、空间记忆、适应性规避），但我们缺乏行为模型。我们扩展GSG框架，通过在线学习而非均衡计算来处理模型未知的策略性生态对手。

野生动物安全中的自适应学习。 以往研究在部署过程中学习对手行为模型：用于巡逻分配的在线算法（Klıma 等人，2014年）、用于对手适应的 SHARP（Kar 等人，2015年）、构建隐式模型的 MINION-sm（Gholami 等人，2019年）以及采用多臂老虎机的 LIZARD（Xu 等人，2021年）。HERDS 则不同，它直接学习防御者策略，而无需构建显式的对手模型，从而具备与模型无关的稳健性。

3. 问题形式化

3.1. 问题概述

我们将人象冲突缓解建模为一个在线组合优化问题。森林-农业边界被划分为 N 个离散区段。防御者在 T 轮次中，将K 名守卫（$K \ll N$）分配到这些区段，以尽量减少大象侵入造成的累计作物损害。对手（大象）根据未知的行为模型选择入口点，而防御者只能获得部分反馈：在有守卫的区段可以观测到拦截情况，但无法得知未守卫区段所使用的具体入口点。目标是学习一种自适应的守卫部署策略，使其遗憾相对于最佳固定事后策略最小。

3.2. 与经典安全博弈的关键差异

我们的建模在三个方面与经典绿色安全博弈存在根本差异，因此需要采用新的算法方法：

挑战 1：对手行为未知。 经典绿色安全博弈假设可以根据历史数据或领域知识估计对手响应模型，从而计算均衡（Fang& Nguyen，2016年）。在人象冲突中，大象行为策略仍然未知。我们观察到大象的策略性行为（最优觅食、自适应规避），但缺乏关于大象如何响应守卫部署的模型。防御者必须在没有任何先验行为假设的情况下学习有效策略。

挑战 2：具有混杂反馈的部分可观测性。 标准安全博弈假设防御者能够观测所有目标上的攻击结果。在人象冲突中，防御者只能观测到设防区段中的拦截情况。大象成功袭击农作物时，总损害是可观测的，但由于监控有限，无法确定它们从哪个具体边界区段进入。这会产生混杂半强盗反馈，也就是说，未设防目标的收益彼此相互依赖，无法单独归因。

挑战 3：非平稳收益。 与目标价值固定的经典情境不同，人象冲突中的收益会随着大象不断调整策略而变化。边界区段的价值函数取决于大象当前的行为，而这种行为会随着大象对已观察到的守卫部署模式作出响应而改变。

表 *1.* 与经典安全博弈的比较

方面	经典广义斯塔克尔伯格博弈	本研究
对手模型	已知/可学习	未知
收益	已知、固定	未知、非平稳
反馈	完全反馈或老虎机反馈	半老虎机，混杂
解概念	均衡	遗憾最小化

这些挑战要求采用一种在线学习方法，该方法能够：**(i)** 在不作行为假设的情况下提供保证；**(ii)** 处理攻击入口无法识别的混杂观测；以及 **(iii)** 通过持续探索适应非平稳收益。

3.3. 正式博弈定义

参与者与策略。 令 $[N] = \{1, \ldots, N\}$ 表示边界区段（目标）的集合。防御者拥有K 资源，最多可保护 K 个目标。防御者的纯策略是一个二进制向量 $\mathbf{v} \in \{0,1\}^N$，其中 $v_i = 1$表示目标 i 受到保护，且 $\|\mathbf{v}\|_1 \leq K$。所有可行纯策略的集合记为 $[V]$。混合策略是定义在 $[V]$上的概率分布。

攻击者可以同时攻击多个区段。令 Q 表示每轮最多可攻击的目标数量。攻击者的纯策略为$\mathbf{a} \in \{0,1\}^N$，其中$a_i = 1$若目标 i 遭到攻击，且 $\|\mathbf{a}\|_1 \leq Q$。攻击者纯策略集合为 $[A]$。

收益与效用。防御者在目标 i 上的收益取决于覆盖情况和攻击状态。令 U_i^c 表示目标 i 被覆盖且遭到攻击（成功拦截）时的收益，U_i^u表示目标未被覆盖且遭到攻击（作物损害）时的收益。由于拦截更可取，因此有 $U_i^u \leq U_i^c$，收益范围限定为 $[-0.5, 0.5]$ 以进行归一化。至关重要的是，这些收益对防御者而言事先未知，必须通过在线学习获得。

在第 t轮，给定策略$\mathbf{v}_t$ 和$\mathbf{a}_t$，防御者的效用为：

$$u(\mathbf{v}_t, \mathbf{a}_t) = \sum_{i \in [N]} v_{t,i} a_{t,i} U_i^c + \sum_{i \in [N]} (1 - v_{t,i}) a_{t,i} U_i^u \quad (1)$$

定义 $r_{t,i} = a_{t,i}[U_i^c - U_i^u] \in [0,1]$和 $C(\mathbf{a_t}) = \sum_{i \in [N]} a_{t,i} U_i^u$后，效用可简化为：

$$u(\mathbf{v}_t, \mathbf{a}_t) = \mathbf{v}_t \cdot \mathbf{r}_t + C(\mathbf{a}_t) \quad (2)$$

博弈协议与目标。 博弈进行 T轮。在每一轮 t：(1) 防御者根据历史记录 F_{t-1}选择 $\mathbf{v}_t \in [V]$；(2) 攻击者选择 $\mathbf{a}_t \in [A]$；(3) 防御者获得效用 $u(\mathbf{v}_t,\mathbf{a}_t)$ 并观察到部分反馈。我们假设 $\mathbf{r}_t$ 可能取决于历史记录 F_{t-1}，但不取决于当前动作 $\mathbf{v}_t$。

目标是最小化累积遗憾：

$$R_T = \max_{\mathbf{v}\in[V]} \sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}_t\right] \tag{3}$$

其中，第一项是最优固定后视策略的效用。遗憾以固定策略为基准，因为当对手可以任意选择任意$\mathbf{a}_t$ 时，最优自适应策略无法实现。

反馈模型。 在每一轮 t中，防御者会观察到：

受守卫区段 ($v_{t,\,i} = 1$)：攻击指示 $\mathbb{I}[a_{t,i} = 1]$和奖励$r_{t,\,i}$（若遭到攻击），以及

未守卫区段($v_{t,\,i} = 0$)：只有作物损害的汇总值 $L_t = \sum_{i:\,v_{t,\,i}=0} a_{t,\,i}|U_i^u|$，无法识别各个入口点。

这构成了混杂半强盗反馈：仅观测到被选臂的逐臂奖励，而未选臂则产生无法归因到具体组成部分的聚合信号。标准半强盗算法假设每个被选动作都能独立观测其奖励；我们的设定不满足这一假设，因此需要一种收益估计机制，将观测到的损失分配给各个候选入口点（第 4.2节）。

4. 战略守卫的资源部署

4.1. 基线：扰动领导者跟随框架

我们采用带均匀探索的扰动领导者跟随算法（FPL-UE）（Xu 等人，2016年）作为基础。该算法用于解决具有半老虎机反馈和未知对手的在线组合优化问题。在介绍我们的改进之前，我们先回顾 FPL-UE 的关键组件。

4.1.1. FPL 算法

扰动领导者跟随（FPL）算法（Neu 与 Bart´ok，2013年）可在半老虎机反馈下解决在线组合优化问题。在每一轮 t中，学习者根据每个目标 i的奖励估计值 $\hat{r}_{t,\,i}$ 从有限动作集合中选择一个动作；这些估计值在没有先验知识的情况下初始化（$\hat{r}_{1,\,i} = 0$）。防御者的纯策略 v_t 计算如下：

$$v_t = \arg\max_{v\in V}\{v \cdot (\hat{\mathbf{r}}_t + \mathbf{z})\}, \tag{4}$$

其中，$\mathbf{z}$ 是一个随机扰动向量，其各分量 $z_i \sim \exp(\eta)$ 独立地服从参数为 η的指数分布。奖励估计值 $\hat{r}_{t+1,\,i}$ 取决于 $p_{t,\,i}$，即选择目标 i的概率。由于 $p_{t,\,i}$ 无法以封闭形式高效计算，带有截断参数 M的几何重采样（GR）算法通过测量重复出现的时间来估计 $p_{t,\,i}$ 从而有效地将 $\frac{1}{p_{t,\,i}}$ 估计为 $P_{(t,i)}$。

4.1.2. FPL-UE：均匀探索组件

FPL-UE（Xu 等人，2016年）通过引入探索扩展了 FPL，以处理$p_{t,\,i}$ 估计值可能任意小的情形。在每一轮次 t，该算法以概率 γ在探索与利用之间进行选择。在探索阶段，从$\mathcal{E}_{\text{expl}} = \{v_1, \ldots, v_N\}$中随机选择一个策略，其中目标 i 在纯策略 v_i中受到保护。因此，每个目标被覆盖的概率为$\frac{\gamma}{N}$。在利用阶段，优化策略 v_t 基于奖励估计值$\hat{\mathbf{r}}_t$ 和随机扰动$\mathbf{z}$进行选择。GR与 FPL-UE算法见附录C。

4.2. 自适应收益学习

HEC 场景（第3.2节）中的一个关键挑战是，边界区域的所有收益并非可独立观测。尽管在已覆盖目标处成功拦截所获得的奖励 U_i^c可以直接观测，但未覆盖目标处的惩罚U_i^u构成了一个根本性的推断问题。当大象成功袭击农作物时，防御者可以观测到农田的损害，但由于有限监控，无法确定其从哪个边界区域进入。由于大象可能通过多个未覆盖边界区域进入同一农田，收益值本质上相互依赖，从而违反了经典安全博弈的独立性假设。

动态收益估计机制为应对收益不确定性，我们设计了一种动态学习机制，在游戏过程中实时更新奖励和惩罚估计。

我们假设目标收益服从未知分布，防御者在游戏过程中获得样本观测。设 MaxInterceptions 表示所有轮次中任一受保护目标拦截的大象轨迹的最大数量，MaxCropRaidLoss 表示任一轮次中观测到的最大作物损害。

对于第 t 轮的目标 i，覆盖奖励 U_i^c 根据拦截的轨迹数量进行估计，并通过 MaxInterceptions 归一化，得到 $U_i^c \in [0, 0.5]$。未覆盖惩罚 U_i^u 根据观测到的作物损害进行估计

在第 t 轮中，通过 MaxCropRaidLoss 归一化，得到 $U_i^u \in [-0.5, 0]$。由于我们无法确定具体使用了哪个未覆盖的边界区域，因此将该轮 t 中作物袭击造成的总损失平均分配给所有未覆盖目标。这种保守的方法确保所有可能易受攻击的目标都计入惩罚估计值，从而满足部分可观测性约束。随后为每个目标计算奖励 $r_{t,i} = a_{t,i}[U_i^c - U_i^u] \in [0, 1]$，并用于在线学习算法中更新奖励估计 $\hat{\mathbf{r}}_t$。

4.3. HERDS算法

4.3.1. 自适应探索-利用预算分配

我们对经典的 FPL-UE 框架进行了修改，以适应人象冲突缓解的独特特点。HERDS 不会像以往那样在每一轮都将所有 K 资源仅分配给探索或利用，而是动态地将资源划分为探索预算 K_{expl} 和利用预算 K_{expt}，其中 $K_{\text{expl}} + K_{\text{expt}} = K$ 和 K_{expl}，$K_{\text{expt}} \in \mathbb{Z}^+$。这种资源划分由一个自适应权衡参数 γ_{t+1} 控制，该参数基于观测到的作物损害，并定义为

$$\gamma_t = \frac{\texttt{CropRaidLoss}_{t-1}}{\texttt{MaxCropRaidLoss}}, \tag{5}$$

其中，分子为第 $t-1$ 轮的作物损害，分母为第 t轮之前观测到的最大损害。

探索预算设为 $K_{\text{expl}} = \lfloor \gamma_t K \rfloor$，利用预算设为 $K_{\text{expt}} = K - K_{\text{expl}}$。

理由： 观察到的作物损害越大，说明策略越无效，因此需要加强探索，以确定最优防护目标。另一方面，较低的损害则表明策略成功，促使我们利用已知的有效目标。此自适应机制是 HEC 场景所特有的，它利用作物损害的观测结果，动态平衡探索与利用。

4.3.2. 目标选择策略

在每一轮 t中，第一步是计算利用目标 $[N_{\text{expt}}]$。具体做法是考虑$[V_{\text{expt}}]$，其中包含所有满足$\|v\|_1 = K_{\text{expt}}$的纯策略 v。随后，对每个 i，从 $[N]$中按$z_i \sim \exp(\eta)$采样扰动$\mathbf{z}$。与 FPL 类似，最优利用策略按如下方式确定：$v_{\text{expt}} = \text{argmax}_{v \in V_{\text{expt}}} \{v \cdot (\hat{\mathbf{r}}_t + \mathbf{z})\}$，由此确定 $[N_{\text{expt}}]$。通过从 $[N] \setminus [N_{\text{expt}}]$中随机抽取$K_{\text{expl}}$ 个目标，形成探索目标集 $[N_{\text{expl}}]$。最后，在所有选定的探索目标和利用目标处部署守卫。

4.3.3. 完整 HERDS 算法

算法1 展示了完整的 HERDS 流程，该流程整合了自适应预算分配、动态收益估计以及经过修改的 FPL-UE 框架。HERDS 的关键创新包括：根据观察到的作物损害进行自适应资源划分（第 4、14 行）；同时进行探索与利用，而非在两种模式之间概率切换（第 7–9 行）；以及能够考虑部分可观测性和目标价值相互依赖关系的动态收益估计（第 12–13 行）。这些改进确保 HERDS 能够针对行为未知、收益不确定且动态演变的自适应生态对手，有效学习最优的守卫部署策略。第 11 行使用标准的几何重采样（GR）算法（算法2，见附录C）。附录5 中的图B.1以可视化方式展示了该算法，便于理解。

算法 1 HERDS算法

要求： $\eta \in \mathbb{R}^+$，$M \in \mathbb{Z}^+$；

1：初始化估计奖励 $\hat{\mathbf{r}} = \mathbf{0} \in \mathbb{R}^N$；2：初始化 $\gamma_1 = 1$；3：对于 $t = 1, \ldots, T$，执行 4：计算 $K_{expl} = \lfloor \gamma_t K \rfloor$ 和 $K_{expt} = K - K_{expl}$；5：找到 $[V_{expt}]$，即防御者所有纯策略的集合，使得 $\|\mathbf{v}\|_1 = K_{expt} \forall \mathbf{v} \in [V_{expt}]$；6：针对 $i \in [N]$，独立抽取 $z_i \sim \exp(\eta)$，并令 $\mathbf{z} = (z_1, \ldots, z_N)$；7：计算 $v_{expt} = \arg\max_{v \in V_{expt}} \{v \cdot (\hat{\mathbf{r}} + \mathbf{z})\}$ 以找到 $[N_{expt}]$；8：通过从 $[N] \setminus [N_{expt}]$中随机抽取 K_{expl} 个目标，选取探索目标 $[N_{expl}]$；9：构造 $v_t \in [0, 1]^N$，使得 $v_t^{(i)} = 1, \forall i \in [N_{expt}] \cup [N_{expl}]$；10：对手选择 $a_t \in [0, 1]^N$，防御者执行 v_t；11：运行 GR($\eta, M, \hat{\mathbf{r}}$ t)：估计 $p_{t,i}$ 为 $P(t, i)$；12：更新 $\hat{\mathbf{r}}(t) \leftarrow \hat{\mathbf{r}}(t) + P(t, i) r_t(i) \cdot \mathbb{I}(t, i)$，其中$\mathbb{I}(t, i) = 1$，对于满足 $v_{t,i} = 1$的 i；否则为 $\mathbb{I}(t, i) = 0$；13：计算 CropRaidLoss t 并更新 MaxCropRaidLoss；14：计算γ_{t+1} = CropRaidLoss t MaxCropRaidLoss；15：循环结束

4.3.4. 性能分析

HERDS 继承了 FPL-UE 的理论基础（Xu 等人，2016年），同时引入了自适应探索。我们推导出以下累积遗憾上界。

$$R_T \le 2T(1-\gamma_t)K\left(1-\frac{\gamma_t}{N}\right)^M + \frac{(1-\gamma_t)K(\log N+1)}{\eta} + \eta m T \min(m,(1-\gamma_t)K) + \sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\right]K\,. \tag{6}$$

与 FPL-UE 的主要区别在于我们采用了自适应探索参数 $\gamma_{t+1} = \texttt{CropRaidLoss}_t \,\overline{\texttt{MaxCropRaidLoss}}$，从而实现目标级采样，而非策略级采样。尽管这种自适应性使分析更加复杂，但当 γ_t 保持有界且不超过某个常数 $\hat{}\gamma$（即 $\gamma_t \le \hat{}\gamma$ 对所有 t 均成立）。完整推导见附录D。

与 FPL-UE 相比，HERDS 通过目标级采样实现了两项关键优势：（1）提高了每个目标的最低覆盖概率，使奖励估计值 $\hat{}r$ 更快趋于稳定；（2）加速了由探索导致的遗憾衰减。如公式 17 （附录 D）所示，HERDS 确保目标被访问的概率不会低于 FPL-UE，从而实现更全面的探索，进而得到更好的奖励估计值。探索效果的提升降低了估计偏差，使 HERDS 能够使用更小的截断参数 M，同时达到与 FPL-UE 相当或更优的性能。该参数控制奖励估计中的探索深度。实验验证了次线性遗憾，并表明算法在不同 M 取值下均具有鲁棒性（图 2）。附录 E介绍了一项消融研究：将 FPL-UE 修改为使用自适应 γ_t。我们发现，这一改动提升了 FPL-UE 的表现，但仍不及 HERDS。

5. 数值实验

5.1. 实验设置

仿真环境。 尽管实地实验能够提供最直接的验证，但人象冲突涉及多个利益相关方，且对人类和野生动物都存在固有风险，因此无法进行直接的实地测试。因此，我们采用空间显式原型基于主体的模型（Purathekandy 等人，2024年），该模型利用印度佩里亚尔-阿加斯蒂亚马莱综合体的大象移动数据进行校准，作为评估 HERDS 的虚拟实验室。该基于主体的模型具有三项关键用途：(i) 提供安全且可重复的测试平台，使策略学习不会造成现实后果；(ii) 通过纳入基于生物学知识的行为模型，绕过野生动物领域常见的严重数据限制；(iii) 能够针对多种对抗策略开展系统评估，而这在实地环境中无法实现。

游戏配置。 研究区域（印度南部西高止山脉的佩里亚尔-阿加斯蒂亚马莱地区）内的连续森林-农业边界被离散化为 $N = 57$ 个边界区段。每场游戏进行 $T = 100$轮次，每轮模拟基于主体的模型中的 10 天大象移动。我们改变防御者的资源预算 $K \in \{3,4,5,6,7,8\}$ 和 GR 截断参数 $M \in \{3,8,15\}$，以评估算法在不同资源约束和估计精度要求下的性能。所有实验中的扰动参数均设为$\eta = 0.5$。

对手模型。 由于 HERDS 与模型无关，我们通过使用两种涵盖不同理性程度的对手模型对其进行评估，以检验其稳健性。这些模型并非嵌入我们算法中的假设，而是用于对学习得到的策略进行压力测试、对大象行为作出合理描述的模型。

（i）短视对手模型（MAM）： 大象智能体没有守卫巡逻位置的记忆或知识，并且对所有边界区段一视同仁。对手无法观测防御者的策略，代表理性程度谱系中最不具策略性的一端。

（ii）有限理性斯塔克尔伯格攻击者模型（BRSAM）： 大象智能体起初不了解守卫位置，但会在游戏过程中通过遭遇逐步建立空间记忆。一旦获悉守卫位置，它们便会自适应地躲避受守卫区段，并利用未受守卫的区域。该模型刻画了没有完美理性的自适应、基于记忆的行为，代表着更具策略性的对手。两种模型的示例轨迹见附录 B.2。MAM 和 BRSAM 模型的完整实现与源代码见补充材料。

基线方法。 我们将 HERDS 与两种基线方法进行比较：FPL-UE（Xu 等人，2016年）和静态策略。在后者中，我们使用基于部分可观测性的基于主体的模型仿真生成固定的守卫部署策略，以保护历史数据中大象最常攻击的边界区段。

评估指标。 我们使用三项不同的指标进行评估：**(i) 累积遗憾** (R_T) 用于衡量最优事后策略的效用与算法效用之间的差距，从而反映学习效率；**(ii)作物侵袭损失**衡量大象突破防御后消耗或损毁的干物质重量（千克），反映现实影响；**(iii) 拦截率** 是指在大象进入农场之前，于设防边界处改变其移动方向的比例，用以量化避免的冲突。

5.2. 结果

我们的实验围绕四个关键研究问题展开，以验证我们的贡献。

研究问题1： 与基线方法相比，*HERDS*能否实现更低的遗憾，并且收敛速度有多快？
研究问题2： 在不同资源预算下，*HERDS*能否有效减少作物损害？
研究问题3： 自适应收益学习机制能否成功识别高价值保护目标？
研究问题4： *HERDS*对不同对手行为模型的鲁棒性如何？

5.2.1. RQ1：学习效率与收敛

图2 展示了在不同资源预算和对手模型下，HERDS、FPL-UE与静态策略在100个博弈轮次中的累积遗憾。我们观察到，在所有实验配置中，HERDS的遗憾始终低于两个基线；表2量化了这一优势，表明与FPL-UE相比，HERDS将最终遗憾降低了15%–45%。除整体性能外，HERDS大约在40–50轮次内达到接近最优的性能，而FPL-UE需要60–80轮次。这种更快的收敛在现实场景中至关重要，因为早期损失会产生显著的社会经济影响。最后，改变截断参数M（这会影响概率估计精度）表明，即使估计值精度较低（$M = 3$），HERDS仍能保持稳定性能，而FPL-UE的性能则有所下降。选择 M 会隐式控制奖励估计中的探索深度，从而影响算法如何适应对手的行为。

5.2.2. RQ2：作物损害减少

遗憾衡量学习效率，而作物损害量化现实世界中的保护影响。图 3展示了两种对手模型在不同资源预算下的平均作物侵袭损失，结果表明，面对配备 $K = 6$ 名守卫的自适应 BRSAM 对手时，HERDS 将平均作物损失从 650 千克降至 350 千克（减少 46%），且比 FPL-UE 优 30%。HERDS 在所有资源预算下均优于基线。尽管所有算法的收益都随资源增加而递减，HERDS 以更少的资源（$K = 6$）即可达到近乎最优的损害减少，而 FPL-UE 需要（$K = 8$），这表明其资源利用效率更高。面对更智能的 BRSAM 对手时，所有方法的作物损失都更高，但 HERDS 仍保持更大的性能优势（减少 20-40%），高于其相对于 MAM 的优势（减少 15-25%），这证明了在面对自适应对手时同时进行探索与利用的价值。

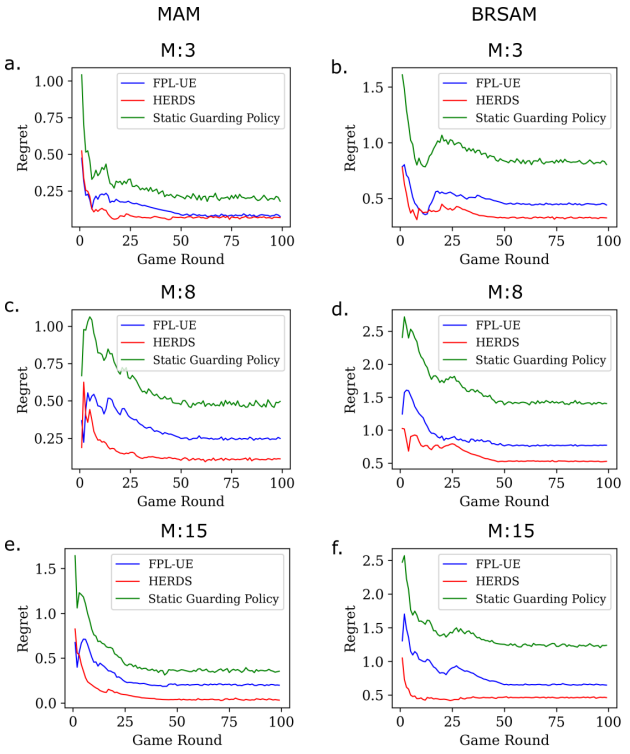

图 *2.* 在不同资源预算和对手模型下，HERDS、FPL-UE 与静态守卫策略在 100 个博弈轮次中的累积遗憾对比。左列展示针对 MAM（短视对手模型）的结果，右列展示针对 BRSAM（有限理性斯塔克尔伯格攻击者模型）的结果，各行表示不同的 GR 截断参数。使用 $K = 5$ 个守卫。

5.2.3. RQ3：自适应收益学习

图4 展示了游戏过程结束时各边界区段的归一化估计奖励$\hat{r}_i$，揭示了 HERDS 学到的空间模式。各边界区段的奖励估计值存在显著差异，归一化数值范围为 0.25 至 1.0，这验证了我们的假设：并非所有边界区域都具有同等的战略重要性。高奖励区段形成了明显的空间集群，分别对应农业区内靠近水源的位置、距离高价值作物较近的位置，以及有利于大象移动的地形特征；这些模式与已知的作物侵袭生态驱动因素相符（Purathekandy 等人，2024年；Shaffer 等人，2019年）。MAM 和 BRSAM 场景下的奖励模式有所不同：在 BRSAM 场景中，随着大象学会避开守卫密集的区域，高奖励区域也随之转移，这表明 HERDS 能够成功地根据不断变化的对手行为调整其收益估计值。我们的收益学习机制（第4.2节）成功应对了部分可观测性约束，即无法直接观测具体入口点；尽管信息不完整，仍能构建准确的奖励估计，方法是

表 2. 针对 MAM 和 BRSAM 对手模型进行游戏对抗结束时的遗憾值。数值越低越好。

K	M=3				M=8				M=15			
	MAM		BRSAM		MAM		BRSAM		MAM		BRSAM	
	FPL-UE	HERDS	FPL-UE	HERDS	FPL-UE	HERDS	FPL-UE	HERDS	FPL-UE	HERDS	FPL-UE	HERDS
3	0.101	0.077	0.222	0.097	0.177	0.079	0.321	0.214	0.157	0.051	0.387	0.314
4	0.111	0.083	0.341	0.153	0.248	0.099	0.411	0.342	0.186	0.055	0.511	0.428
5	0.118	0.101	0.487	0.245	0.268	0.116	0.671	0.528	0.213	0.078	0.673	0.518
6	0.243	0.122	0.499	0.269	0.311	0.132	0.677	0.574	0.342	0.087	0.741	0.588
7	0.255	0.146	0.542	0.321	0.387	0.138	0.692	0.592	0.369	0.089	0.789	0.594
8	0.287	0.148	0.572	0.375	0.422	0.156	0.748	0.621	0.412	0.100	0.841	0.621

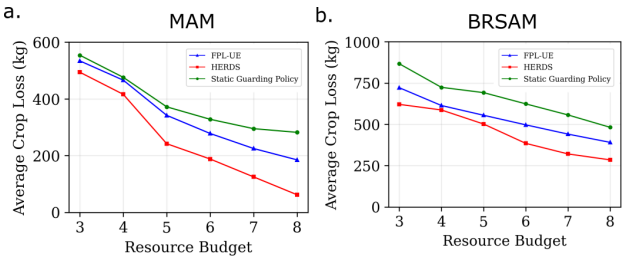

图3. 在 $M = 8$条件下，(a) MAM 和 (b) BRSAM 对手模型的平均作物侵袭损失（干物质千克数）随守卫资源预算的变化。

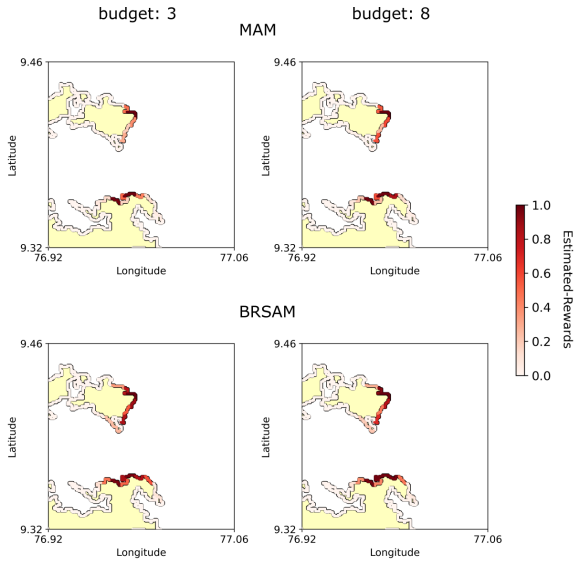

图4. 100 轮游戏过程结束时边界区段归一化估计奖励的空间分布，对比不同资源预算和对手模型的情形（上排为 MAM，下排为 BRSAM）。颜色强度表示奖励大小，较暖的颜色表示更高的战略价值。

将惩罚分配到未覆盖路段。

5.2.4. 研究问题4：对手模型鲁棒性

分析算法对抗 MAM 和 BRSAM 对手时的性能，有助于了解算法的稳健性。在两种模型中，HERDS 的表现均优于基线方法，凸显了其应对攻击者行为不确定性的韧性。面对更强的 BRSAM 对手，HERDS 相较于 FPL-UE 的优势进一步扩大，表明同时采用探索-利用策略对抗自适应对手具有重要价值。图3 显示，静态策略在对抗 BRSAM 时表现不佳（作物损失为 500–800 千克），因为适应性强的大象很快就会学会绕过固定的守卫位置；而 HERDS 通过持续更新部署策略，仍能保持有效（作物损失为 250–600 千克）。

6. 结论

本文提出了首个在线博弈论学习框架，用于在行为未知且可观测性有限的情况下，对自适应生态对手进行战略防御。我们提出了 HERDS，该算法通过动态探索–利用划分、在混杂反馈下进行自适应收益估计以及联合资源部署，对 FPL-UE 进行了扩展。该算法无需依赖行为假设即可实现次线性遗憾；当对手模型不可用时，这一特性尤为关键。

使用经大象移动数据校准的基于主体的模型进行实验验证，得出了三项关键发现：（1）基于损失的预算分配，与均匀探索相比，可将遗憾降低 15–45%，这表明已观察到的损害有助于制定探索策略。（2）将未观察到的损失分配到候选入口点，即使存在混杂的半老虎机反馈，也能成功学习空间模式，且学习得到的聚类与生态驱动因素相符。（3）在不同对手理性水平下均表现稳健，作物损害减少 40–50%，收敛速度加快 30–50 %（40–50 轮次，而非 60–80 轮次）。

除了人象冲突之外，我们的框架还可推广到存在策略性但未建模对手的安全领域，例如反偷猎巡逻、基础设施保护和竞争性资源分配。对于保护工作，HERDS 提供了可执行的部署策略，且仅需部分可观测性。未来的工作应验证其实地表现，并纳入守卫效能不确定性。

影响声明

本研究运用安全博弈理论和在线机器学习的视角分析人象冲突，并据此提出一种有助于推动机器学习这一更广泛领域发展的算法。尽管我们的框架有望支持野生动物保护和安全工作，但也可能被滥用于侵入式野生动物监控或强制控制系统。冲突缓解的首要目标是减少对当地社区和脆弱大象种群造成的伤害，因此将野生动物描绘成“对手”可能存在问题，因为这有可能过度简化复杂的社会—生态关系。任何现实世界中的部署都必须征得受影响社区的同意，并通过广泛的利益相关方对话，认真关注动物福利。

参考文献

An, B., Pita, J., Shieh, E., Tambe, M., Kiekintveld, C. 和 Marecki, J. GUARDS 和 PROTECT：安全博弈的新一代应用。*SIGecom Exch.*，10(1)：31–34，2011年3月。doi：10.1145/1978721.1978729。

Audibert, J.-Y. 和 Bubeck, S. 部分监控下的遗憾界与极小极大策略。机器学习研究杂志，11：2785–2836，2010年。

Auer, P., Cesa-Bianchi, N. 和 Fischer, P. 多臂老虎机问题的有限时间分析。机器学习，47(2–3)：235–256，2002年。doi：10.1023/A: 1013689704352。

Awerbuch, B. 和 Kleinberg, R. D. 在线线性优化与自适应路由。计算机与系统科学杂志，74(1):97–114，2008年。doi: 10.1016/j.jcss. 2007.04.016。

Cesa-Bianchi, N. 和 Lugosi, G. 预测、学习与博弈。剑桥大学出版社，剑桥，2006年。ISBN 978-0-521-84108-5。

Fang, F. 和 Nguyen, T. H. 绿色安全博弈：运用博弈论应对绿色安全挑战。*SIGecomExchanges*，15(1):78–83，2016年。doi: 10.1145/ 2994501.2994507。

Fang, F.、Stone, P. 和 Tambe, M. 当安全博弈走向绿色：设计防御者策略以防止偷猎和非法捕捞。载于第二十四届国际人工智能会议论文集，IJCAI’15，第2589–2595页。AAAI出版社，2015年。ISBN 9781577357384。

Fernando, P., Kumar, M. A., Williams, A. C., Wikramanayake, E., Aziz, T., 和 Singh, S. M. 《南亚人象冲突缓解措施综述》。世界自然基金会 - 世界自然基金会，2008年。

Frank, D. M. 和 Sarkar, S.《生物多样性保护中的群体决策：博弈论的启示》。*PLOSONE*，5(5):1–10，2010年05月。doi: 10.1371/journal.pone. 0010688。

Gholami, S., Yadav, A., Tran-Thanh, L., Dilkina, B., 和 Tambe, M.《不要把所有策略都放在一个篮子里：利用不完美先验知识参与绿色安全博弈》。载于第*18*届自主智能体与多智能体系统国际会议论文集，AAMAS ’19，第395–403页。自主智能体与多智能体系统国际基金会，2019年。ISBN 9781450363099。

Janssen, S., Matias, D., 和 Sharpanskykh, A. 一种用于机场安全巡逻的基于智能体的经验博弈论方法。*Aerospace*，7(1)，2020年。ISSN 2226- 4310。doi: 10.3390/aerospace7010008。

Kalai, A. T. 和 Vempala, S. S. 用于在线决策问题的高效算法。计算机与系统科学杂志，71(3):291–307，2005年。doi: 10.1016/j.jcss.2004. 10.016。

Kar, D., Fang, F., Delle Fave, F., Sintov, N. 和 Tambe, M.《权力的游戏：当人类行为模型在重复斯塔克尔伯格安全博弈中相互竞争》。载于*2015*年国际自主智能体与多智能体系统会议论文集，AAMAS ’15，第1381–1390页。国际自主智能体与多智能体系统基金会，2015年。ISBN 9781450334136。

Kirkland, L.-A.、de Waal, A. 和 deVilliers, J. P. 在地理空间框架下评估用于野生动物安全的纯策略斯塔克尔伯格博弈。载于 Gerber, A.（编），人工智能研究，第 101–118 页，沙姆，2020年。施普林格国际出版社。ISBN 978-3-030- 66151-9。

Kl´ıma, R.、Kiekintveld, C. 和 Lis´y, V. 边境巡逻资源分配的在线学习方法。载于 Poovendran, R. 和 Saad, W.（编），安全决策与博弈论，第 340–349 页，沙姆，2014年。施普林格国际出版社。ISBN 978-3-319-12601-2。

Marecki, J.、Tesauro, G. 和 Segal, R. 与未知对手进行重复斯塔克尔伯格博弈。载于*2014*年国际自主智能体与多智能体系统会议论文集，AAMAS ’12，第 821–828 页。国际自主智能体与多智能体系统基金会，2012。ISBN 978-0- 9817381-2-3。doi: 10.5555/2343776.2343814。

McLane, A. J.、Semeniuk, C.、McDermid, G. J. 和 Marceau, D. J. 基于主体的模型在

野生动物生态学与管理中的作用。*Ecological Mod-elling*，222(8)：1544–1556，2011年。ISSN 0304-3800。doi：https://doi.org/10.1016/j.ecolmodel.2011.01.020。

McMahan, H. B. 和 Blum, A. 在老虎机设定下针对自适应对手的在线几何优化。载于第*17*届年度学习理论会议（*COLT*）论文集，计算机科学讲义第3120卷，第109–123页。Springer，2004年。doi：10.1007/978-3-540-27819-1_8。

Neu, G. 和 Bart´ok, G. 一种基于半老虎机反馈的高效学习算法。载于 Jain, S.、Munos, R.、Stephan, F. 和 Zeugmann, T.（编），算法学习理论，第234–248页。Springer Berlin Heidelberg，2013年。ISBN 978-3-642-40935-6。

Paruchuri, P.、Pearce, J. P.、Tambe, M.、Ordonez, F. 和 Kraus, S. 一种针对多个对手的高效安全启发式方法。载于第*6*届自主智能体与多智能体系统国际联合会议论文集，AAMAS ’07。计算机协会，2007年。ISBN 9788190426275。doi：10.1145/1329125.1329344。Purathekandy, A.、Oommen, M. A.、Wikelski, M. 和 Sub-ramani, D. N. 印度佩里亚尔–阿加斯蒂亚马莱复合区大象袭扰农作物动态的基于智能体的模型。生态建模，496：110843，2024年。ISSN 0304-3800。doi：https://doi.org/10.1016/j.ecolmodel. 2024.110843。Qian, Y.、Haskell, W. B.、Jiang, A. X. 和 Tambe, M. 资源保护博弈中保护者最优策略的在线规划。载于*2014*年自主智能体与多智能体系统国际会议论文集，AAMAS ’14，第733–740页，南卡罗来纳州里奇兰，2014年。国际自主智能体与多智能体系统基金会。ISBN 9781450327381。Shaffer, L. J.、Khadka, K. K.、Van Den Hoek, J. 和 Naithani, K. J. 人象冲突：当前管理策略综述与未来方向。《生态学与进化前沿》，第6卷——2018年，2019年。ISSN 2296-701X。doi：10.3389/fevo.2018.00235。Tambe, M. 安全与博弈论：算法、部署系统与经验教训。剑桥大学出版社，2011年。Xu, H.、Tran-Thanh, L. 和 Jennings, N. R. 在没有先验知识的情况下参与重复安全博弈。载于*2016*年自主智能体与多智能体系统国际会议论文集，AAMAS ’16，第104–112页。国际自主智能体与多智能体系统基金会，2016年。ISBN 9781450342391。

Xu, L., Bondi, E., Fang, F., Perrault, A., Wang, K. 和 Tambe, M. 双重使命巡逻：用于绿色安全的多臂老虎机。*AAAI* 人工智能会议论文集，35(17):14974–14982，2021年5月。doi：10.1609/aaai.v35i17.17757。Yang, R., Ford, B., Tambe, M. 和 Lemieux, A. 通过自适应资源分配保护野生动物免受非法盗猎者侵害。载于*2014*年国际自主智能体与多智能体系统会议论文集，AAMAS ’14，第453–460页。国际自主智能体与多智能体系统基金会，2014年。ISBN 9781450327381。Zhu, J., Zhang, J., Ling, Q. 和 Dullerud, G. E. 在无先验知识且可观测性有限的情况下，适用于重复安全博弈的低资源再分配防御策略。*IEEE*认知与发展系统汇刊，15(4):2156–2168，2023年。doi: 10.1109/TCDS.2023.3241364。

A. 符号与记法

表 3 总结了论文中使用的符号和记法。

表 3. 符号汇总

符号	描述
博弈设定	
N	边界区段（目标）数量
K	可部署守卫资源数量
T	博弈轮次总数
Q	大象每轮可瞄准的边界区段最大数量
ΔL	每个边界区段的固定长度
$[N]$	所有目标的集合 $\{1, 2, \ldots, N\}$
策略	
$\mathbf{v} \in \{0,1\}^N$	防御者的纯策略（二进制向量）
V	防御者的所有纯策略集合
$\mathbf{a} \in \{0,1\}^N$	攻击者的纯策略（二进制向量）
A	攻击者的所有纯策略集合
$\mathbf{v}_t$	第 t 轮防御者的策略
$\mathbf{a}_t$	第 t 轮攻击者的策略
$v_{t,i}$	二元指示变量：若目标 i 在第 t 轮受到保护，则为 1，否则为 0
$a_{t,i}$	二元指示变量：若目标 i 在第 t 轮受到攻击，则为 1，否则为 0
收益与效用	
U_i^c	覆盖（保护）目标 i的收益
U_i^u	未覆盖目标 i的收益
$u(\mathbf{v}_t, \mathbf{a}_t)$	第 t 轮次防御者的效用
$r_{t,i}$	$a_{t,i}[U_i^c - U_i^u]$
$\hat{r}_{t,i}$	目标 i 在第 t 轮次的估计奖励
$\hat{\mathbf{r}}_t$	第 t轮的估计奖励向量
$C(\mathbf{a}_t)$	效用函数中的常数项，$C(\mathbf{a}_t) = \sum_{i \in [N]} a_{t,i} U_i^u$
博弈历史与学习	
F_t	截至第 t轮（含）的完整博弈历史
F_0	无初始历史信息
R_T	第 T轮的遗憾估计
算法参数	
η	指数分布的扰动参数
$\mathbf{z}$	随机扰动向量
z_i	目标 i 的扰动，其中 $z_i \sim \exp(\eta)$
M	几何重采样（GR）算法的截断参数
$P(t, i)$	^{1}的估计值$\overline{p_{t,i}}$ 来自 GR 算法
$p_{t,i}$	第 t轮选择目标 i的概率
γ_t	第 t 轮的探索-利用权衡参数
K_{expl}	探索资源预算
K_{expt}	利用资源预算
$[N_{\text{expt}}]$	为利用选择的目标集合
$[N_{\text{expl}}]$	为探索选择的目标集合
$[V_{\text{expt}}]$	所有具有 $\|\mathbf{v}\|_1 = K_{\text{expt}}$的纯策略集合
E	探索策略集合
$\mathbb{I}(t, i)$	示性函数：若 v_t，则为 $\mathbf{1}_{\ ,i} = 1$，否则为 0
性能指标	
最大拦截次数	在任一目标处拦截的大象轨迹的最大数量
MaxCropRaidLoss	任一轮次中观测到的最大作物损害（kg）
CropRaidLoss_t	第 t 轮次的作物侵袭损失（干物质 kg）

B. 附加图表

B.1. HERDS 工作流程示意图

图中展示了所提出的 HERDS算法的各个组成部分5。

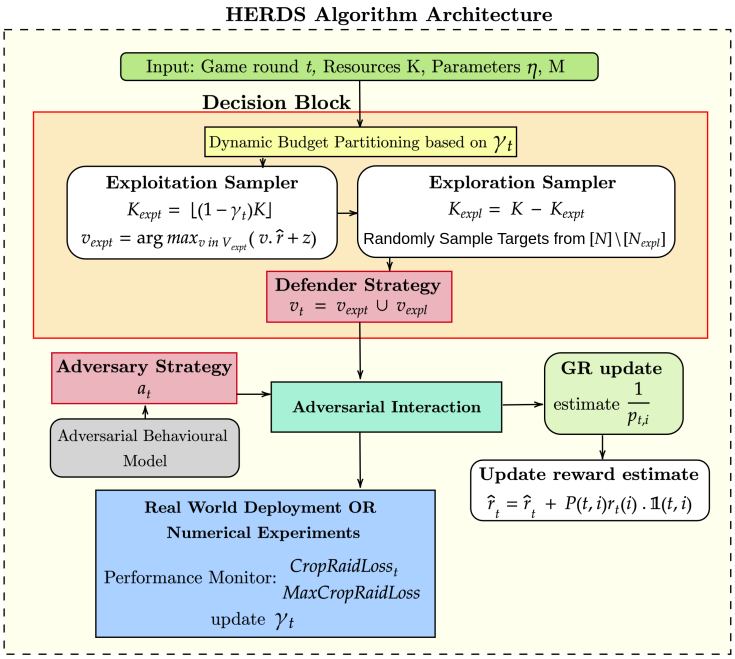

图5. 展示我们HERDS算法各组成部分的流程图。

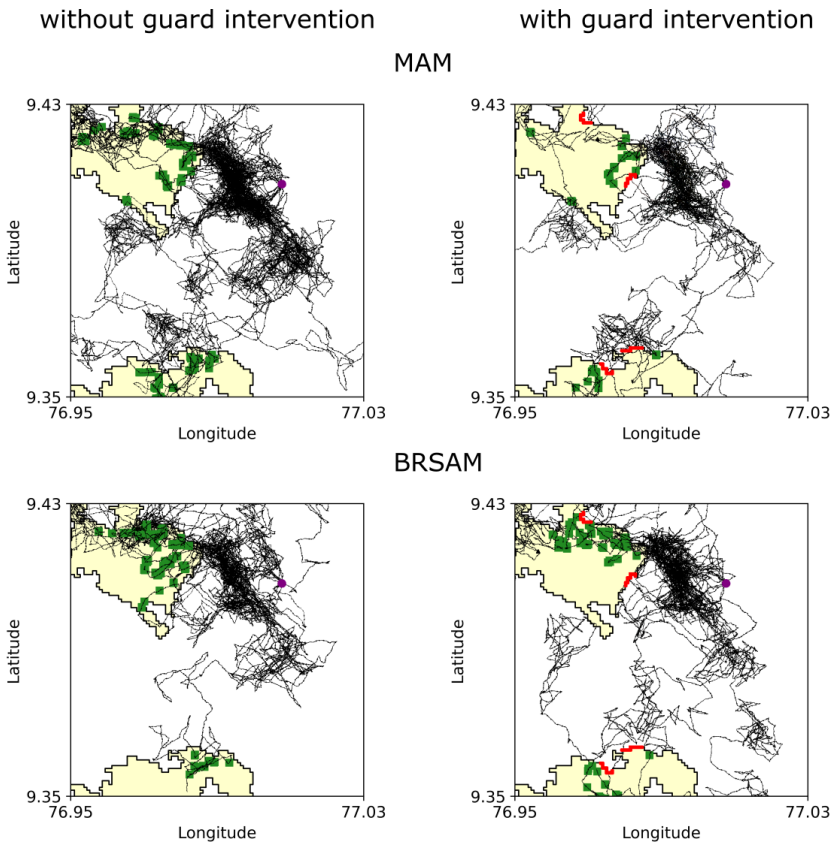

图6. 基于智能体的仿真比较了两种对手模型在无护卫干预（左）与有护卫干预（右）情况下的大象轨迹：短视对手模型（MAM，上）和有限理性斯塔克尔伯格攻击者模型（BRSAM，下）。紫色圆圈标示大象起始位置，黑线表示轨迹，黄色区域为人类聚居地，白色区域为森林，绿色方块为遭大象侵袭的农田，红色斑块为设防边界段。

B.2. 攻击者模型的轨迹模拟

图 6 展示了本文使用的两种攻击者模型——短视对手模型（MAM）和有限理性斯塔克尔伯格攻击者模型（BRSAM）的轨迹演化示例。空间模式展示了自适应对手（BRSAM）如何根据守卫部署调整其接近路线，这也说明了动态防御策略的必要性。这些行为模型确保守卫部署模型能够应对各种潜在对手，并具备稳健性和泛化能力，同时考虑到野外大象实际决策方式的不确定性。

C. GR 和 FPL-UE 算法

算法 2列出了 GR 算法，算法 3 列出了 FPL-UE 算法。这些算法用于本文主文算法 1中所列的 HERDS 算法。

算法 2GR 算法

要求： $\eta \in \mathbb{R}^+, M \in \mathbb{Z}^+, \hat{\mathbf{r}} \in \mathbb{R}^n, t \in \mathbb{N}$；**确保：** $P(t) := \{P(t,1), \ldots, P(t,n)\} \in \mathbb{Z}^N$**1：初始化** $\forall i \in [N]$：$P(t,i) = 0, p = 1$；**2：对于** $p = 1, \ldots, M$，**执行 3：在算法3中使用** γ_t **重复步骤4至9一次，以生成** $\hat{v}$，**作为** v_t**的模拟；4：对于所有** $i \in [N]$，**执行 5：如果** $p < M$ **且** $\hat{v}_i = 1$ **且** $P(t,i) = 0$，**则 6：** $P(t,i) = p$；**7：否则，如果** $p = M$ **且** $P(t,i) = 0$，**则 8：设置** $P(t,i) = M$；**9：结束如果 10：结束循环 11：如果对于所有** $i \in [N]$，$P(t,i) > 0$，**则 12：中断；13：结束如果 14：结束循环**

算法3FPL-UE算法

要求： $\eta \in \mathbb{R}^+$、$M \in \mathbb{Z}^+$、$\gamma \in [0,1]$；**1：初始化估计奖励** $\hat{\mathbf{r}} = \mathbf{0} \in \mathbb{R}^N$；**2：选取探索策略集合** $E = \{v_1, \ldots, v_n\}$，**使得目标** i **在纯策略** v_i**中受到保护；3：对于** $t = 1$，**.....，** T**执行 4：以概率** γ **采样** $flag \in [0,1]$，**使得** $flag = 0$；**5：如果** $flag = 0$ **则 6：令** v_t **为从** E**中均匀随机采样的策略；7：否则 8：对** $i \in [N]$ **独立抽取** $z_i \sim \exp(\eta)$，**并令 z** $= (z_1, \ldots, z_n)$；**9：** $v_t = \arg\max_{v \in V} \{v \cdot (\hat{\mathbf{r}} + \mathbf{z})\}$；**10：结束如果 11：对手选择** $a_t \in [0,1]^N$，**防御者执行** v_t；**12：运行 GR(**$\eta, M, \hat{\mathbf{r}}$ t**)：将** p_t、$_i$ **估计为** $P(t,i)$；**13：更新** $\hat{\mathbf{r}}(t) \leftarrow \hat{\mathbf{r}}(t) + P(t,i) r_t(i) \cdot \mathbb{I}(t,i)$，**其中当** i **满足** v_t、$_i = 1$ **时** $\mathbb{I}(t,i) = 1$；**否则** $\mathbb{I}(t,i) = 0$；**14：结束循环**

D. HERDS 的遗憾界

D.1. 预备知识

防御者在轮次 t的效用为:

$$u(\mathbf{v}_t, \mathbf{a}_t) = \sum_{i\in[N]} v_{t,i}a_{t,i}U_i^c + \sum_{i\in[N]} (1 - v_{t,i})a_{t,i}U_i^u \tag{7}$$

这里， $[N] = \{1, \ldots, N\}$表示边界区段（目标）的集合， $\mathbf{v}_t$表示 t时刻的防御者策略， $\mathbf{a}_t$表示 t时刻的攻击者策略， U_i^c表示覆盖目标 i的收益， U_i^u表示未覆盖目标 i的收益。我们还假设$U_i^u \le U_i^c$，因为覆盖目标总是优于不覆盖目标；同时，为进行归一化，收益有界（U_i^c， $U_i^u \in [-0.5, 0.5]$）。

简化后,

$$u(\mathbf{v}_t, \mathbf{a}_t) = \sum_{i\in[N]} v_{t,i}a_{t,i}(U_i^c - U_i^u) + \sum_{i\in[N]} a_{t,i}U_i^u \tag{8}$$

定义 $r_{t,i} = a_{t,i}[U_i^c - U_i^u] \in [0, 1]$和 $C(\mathbf{a}_t) = \sum_{i\in[N]} a_{t,i}U_i^u$，效用可简化为:

$$u(\mathbf{v}_t, \mathbf{a}_t) = \mathbf{v}_t \cdot \mathbf{r}_t + C(\mathbf{a}_t) \tag{9}$$

步骤 T 的遗憾 R_T 定义为最佳固定事后策略的总累积效用与所用算法的预期累积效用之差:

$$R_T = \max_{\mathbf{v}\in[V]} \sum_{t=1}^{T} u(\mathbf{v}, \mathbf{a}_t) - \mathbb{E}\left[\sum_{t=1}^{T} u(\mathbf{v}_t, \mathbf{a}_t)\right] \tag{10}$$

$$R_T = \max_{\mathbf{v}\in[V]} \sum_{t=1}^{T} (\mathbf{v} \cdot \mathbf{r}_t + C(\mathbf{a}_t)) - \mathbb{E}\left[\sum_{t=1}^{T} (\mathbf{v}_t \cdot \mathbf{r}_t + C(\mathbf{a}_t))\right] \tag{11}$$

$$R_T = \max_{\mathbf{v}\in[V]} \sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}_t\right] \tag{12}$$

最优固定事后策略 $v*$ 定义为

$$v* = \arg\max_{\mathbf{v}\in[V]} \sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v} \tag{13}$$

D.2. 引理

以下引理摘自 Xu 等人（2016年）:

- GR 算法生成估计值 $\hat{r}_{t,\ i}$ 使得

$$\mathbb{E}[\hat{r}_{t,i} \mid \mathcal{F}_{t-1}] = \big(1 - (1 - p_{t,i})^M\big) r_{t,i} \tag{14}$$

其中， $\mathcal{F}_{t-1}$ 是截至时间 $t-1$ 的历史记录， $p_{t,\ i} = \Pr[v_{t,i} = 1]$ 是目标 i 在时间 t受到保护的概率， M 是 GR 算法中的截断参数，$r_{t,\ i}$ 是目标 i 在轮次 t的真实奖励，而 $\hat{r}_{t,\ i}$ 是对目标 i 在轮次 t的奖励估计。

这里，

$$\hat{r}_{t,i} \leftarrow \hat{r}_{t,i} + P_{t,i} r_{t,i} \mathbb{I}(t,i) = 1 \tag{15}$$

其中，$P_{t,\ i}$ 是 $p^{1}_{t\ ,\ i}$ 的估计值，而 $\mathbb{I}(t,\ i)$ 是示性函数：若目标 i 在第 t 轮被防御者选中，则取值为 1；否则取值为 0。

在 FPL-UE 中，算法以概率 γ 进行探索，以概率 $1-\gamma$ 进行利用。由于探索策略集合中的每个目标至少受到一种纯策略的保护，因此 $p_{t,\ i}$ 有如下下界：

$$p_{t,i} \geq \frac{\gamma}{n} \tag{16}$$

在 HERDS 中，算法在每一轮内对部分目标进行探索，同时对其他目标进行利用，并将保护概率 $p_{t,\ i}$ 维持在如下下界之上：

$$p_{t,i} \geq \frac{\gamma_t}{N - K_{expt}} \geq \frac{\gamma_t}{N - (1-\gamma_t)K} \tag{17}$$

其中，γ_t 是第 t 轮的探索-利用权衡参数，K 是可部署守卫资源总数。

这个下界确保即使是根据当前历史记录看起来较差的目标，也会偶尔被抽样，这对于维持无偏奖励估计值是必要的。

- 此外，对于任意轮次 t，奖励向量 r_t 满足

$$\|r_t\|_1 \leq m \tag{18}$$

其中，$r_{t,i} = a_{t,i}[U_i^c - U_i^u]$，且 $U_i^c, U_i^u \in [-0.5, 0.5]$，$a_t \in \{0,1\}^N$ 是攻击者在轮次 t 的纯策略，而 m 是攻击者可同时攻击的最大目标数量（$\|a_t\|_1 \leq m$）。
在HERDS中，$r_{t,i} \in [0,1]$。因此，$\|r_{t,i}\|_1 \leq N$。但由于对攻击者移动的约束，我们可以假设 $\|r_{t,i}\|_1 \leq m$，其中 $m << N$。在最坏情况下，$m = N$。

D.3. 遗憾分解

与先前文献（Xu 等人，2016年）一样，我们定义一些记号来分析HERDS算法。令 $\tilde{v_t}$ 表示若HERDS仅使用全部 K_{expt} 资源进行利用时将选择的策略（$K_{expt} = \lfloor (1-\gamma_t)K \rfloor$）（即 $\tilde{v_t} = v_{expt}$）。
给定最优的固定事后策略 $v*$（公式 13），我们可以将公式 12 改写为

$$R_T = \sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}* - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}_t\right] \tag{19}$$

我们在公式 19中加上再减去 $\sum_{t=1}^{T} \mathbf{r}_t \cdot \tilde{v_t}$

$$R_T = \left(\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}* - \sum_{t=1}^{T} \mathbf{r}_t \cdot \tilde{v_t}\right) + \left(\sum_{t=1}^{T} \mathbf{r}_t \cdot \tilde{v_t} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}_t\right]\right) \tag{20}$$

因此，我们可以将遗憾 R_t 分解为两个部分：

$$R_T = R_T^a + R_T^b \tag{21}$$

让我们考虑公式中的每个单独组成部分。21：

1.

$$R_T^a = \sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}* - \sum_{t=1}^{T} \mathbf{r}_t \cdot \tilde{v_t} \tag{22}$$

这里，使用基于 FPL 的算法选择 $K_{expt} = \lfloor (1-\gamma_t)K \rfloor$ 资源，其中 K_{expt} 随时间步 t变化。

对于 FPL 算法，利用部分的遗憾上界如下所示（Xu 等人，2016年）：

$$R_T^{FPL} \leq 2Tk(1-\frac{\gamma}{N})^M + \frac{k(\log N+1)}{\eta} + \eta m T min(m,k) \tag{23}$$

在 HERDS 中，$k = K_{expt}$ 和 $K_{expt} \leq K$。此外，$\gamma = \gamma_t$。

$$K_{expt} = \lfloor (1-\gamma_t)K \rfloor \approx (1-\gamma_t)K$$

$$R_T^a \leq 2TK_{expt}(1-\frac{\gamma_t}{N})^M + \frac{K_{expt}(\log N+1)}{\eta} + \eta m T min(m, K_{expt}) \tag{24}$$

$$R_T^a \leq 2T(1-\gamma_t)K(1-\frac{\gamma_t}{N})^M + \frac{(1-\gamma_t)K(\log N+1)}{\eta} + \eta m T min(m, (1-\gamma_t)K) \tag{25}$$

2.

$$R_T^b = \sum_{t=1}^{T} \mathbf{r}_t \cdot \tilde{v_t} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t \cdot \mathbf{v}_t\right] \tag{26}$$

这里，$K_{expl} = \lfloor \gamma_t K \rfloor$ 资源是从 N 中随机选取的，并非最优选择。
每个探索资源所产生的奖励最多可能比其替代的利用资源少 1（$r_{t,\ i} \in [0,1]$）。因此，每轮的损失 $\leq K_{expl}$

$$R_T^b \leq \mathbb{E}\left[\sum_{t=1}^{T} K_{expl}\right] \tag{27}$$

$$R_T^b \leq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t K\right] \tag{28}$$

$$R_T^b \leq \sum_{t=1}^{T} \mathbb{E}\left[\gamma_t\right] K \tag{29}$$

D.4. 总遗憾界

综合以上结果，我们得到 HERDS 的遗憾上界 R_T：

$$R_T \leq 2T(1-\gamma_t)K(1-\frac{\gamma_t}{N})^M + \frac{(1-\gamma_t)K(\log N+1)}{\eta} + \eta m T min(m,(1-\gamma_t)K) + \sum_{t=1}^{T} \mathbb{E}\left[\gamma_t\right] K \tag{30}$$

公式中的前三项6 与 Xu 等人（2016年）中的利用遗憾项非常相似。这是意料之中的，因为 HERDS 是采用自适应探索的 FPL-UE 变体。自适应部分 γ_t 增加了分析的复杂性，但在最坏情况分析中，如果能将其界定为 $\hat{}\gamma\,(\gamma \leq \hat{}\gamma)$，我们就能确保 HERDS 算法具有次线性遗憾。

通过取 $\gamma_t \approx \mathcal{O}(\frac{1}{\sqrt{(T)}})$，我们可以得到 HERDS 的遗憾上界为 $\mathcal{O}\sqrt{(T)}$。

E. 消融研究

为了通过实证验证我们提出的框架的结构完整性以及 HERDS算法中各组件的作用，我们通过修改 FPL-UE算法，使其采用基于观测到的农作物袭击损失的自适应 γ_t (公式5)，从而开展了消融研究（FPL-UE-A）。这项消融研究的结果汇总于表4。由于探索能力有所提高，与 FPL-UE算法相比，FPL-UE-A 的遗憾值下降得更快。然而，FPL-UE-A 仍采用策略级采样，即使使用自适应 γ_t，其最终遗憾值也与 FPL-UE 相近。这表明 HERDS 的优势在于目标级采样，而非策略级采样；探索能力的提升能够加快遗憾最小化，而目标级采样则能进一步降低遗憾。

表*4*。 FPL-UE 中采用自适应探索-利用的消融研究（FPL-UE-A）

K	M=3				M=8				M=15			
	MAM		BRSAM		MAM		BRSAM		MAM		BRSAM	
	FPL-UE-A	HERDS	FPL-UE-A	HERDS	FPL-UE-A	HERDS	FPL-UE-A	HERDS	FPL-UE-A	HERDS	FPL-UE-A	HERDS
3	0.975	0.077	0.241	0.097	0.178	0.079	0.375	0.214	0.135	0.051	0.385	0.314
4	0.988	0.083	0.347	0.153	0.242	0.099	0.417	0.342	0.179	0.055	0.524	0.428
5	0.110	0.101	0.452	0.245	0.271	0.116	0.665	0.528	0.220	0.078	0.672	0.518
6	0.237	0.122	0.457	0.269	0.307	0.132	0.687	0.574	0.351	0.087	0.735	0.588
7	0.248	0.146	0.555	0.321	0.351	0.138	0.674	0.592	0.342	0.089	0.745	0.594
8	0.254	0.148	0.561	0.375	0.434	0.156	0.757	0.621	0.401	0.100	0.801	0.621