深度强化学习攻击方法介绍

2022.5.20

攻击目的：

- 降低目标智能体的长期预期奖励：

- 让目标智能体执行攻击者期望的动作以转移到特定的状态：

攻击场景分类：

- 按攻击主要实施阶段进行划分：
  - 训练阶段的攻击
  - 测试阶段的攻击
- 按攻击者对于目标智能体的了解程度进行划分：
  - 白盒攻击
  - 黑盒攻击
- 按攻击者在测试阶段对目标智能体的控制权限进行划分：
  - 高权限的攻击（攻击者可任意修改智能体输入）
  - 受限的攻击（攻击者只能修改智能体的部分输入）

各攻击方法的威胁模型

攻击方法(year, Citations)	攻击目的	实施阶段	最低了解程度	测试阶段权限
FGSM(2017,551)	降低期望	测试阶段	黑盒	高权限
战略时间攻击(2019,271)	降低期望	测试阶段	白盒	高权限
迷惑攻击(2019,271)	转移到指定状态	测试阶段	黑盒	高权限
策略诱导攻击(2017,201)	转移到指定状态	测试阶段	黑盒	高权限
木马攻击(2019,20)	降低期望 & 转移到指定状态	训练阶段	白盒	低权限
对抗性策略(2020,175)	降低期望	测试阶段	黑盒	低权限
2022		?		

FGSM : Fast Gradient Sign Method

FGSM focuses on adversarial perturbations where each pixel of the input image is changed by no more than ϵ. Given a linear function $g(x) = w^\top x$, the optimal adversarial perturbation η that satisfies $\|\eta\|_\infty < \epsilon$ is

$$\eta = \epsilon \operatorname{sign}(w), \tag{1}$$

since this perturbation maximizes the change in output for the adversarial example $\tilde{x}$, $g(\tilde{x}) = w^\top x + w^\top \eta$.

Given an image classification network with parameters θ and loss $J(\theta, x, y)$, where x is an image and y is a distribution over all possible class labels, linearizing the loss function around the input x results in a perturbation of

$$\eta = \epsilon \boxed{\operatorname{sign}}(\boxed{\nabla_x} J(\theta, x, y)). \tag{2}$$

$$\eta = \begin{cases} \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y)) & \text{for constraint } \|\eta\|_\infty \leq \epsilon \\ \epsilon \sqrt{d} * \frac{\nabla_x J(\theta, x, y)}{\|\nabla_x J(\theta, x, y)\|_2} & \text{for constraint } \|\eta\|_2 \leq \|\epsilon \mathbf{1}_d\|_2 \\ \text{maximally perturb highest-impact dimensions with budget } \epsilon d & \\ & \text{for constraint } \|\eta\|_1 \leq \|\epsilon \mathbf{1}_d\|_1 \end{cases}$$

战略时间攻击

$$c(s_t) = \max_{a_t} \pi(s_t, a_t) - \min_{a_t} \pi(s_t, a_t).$$

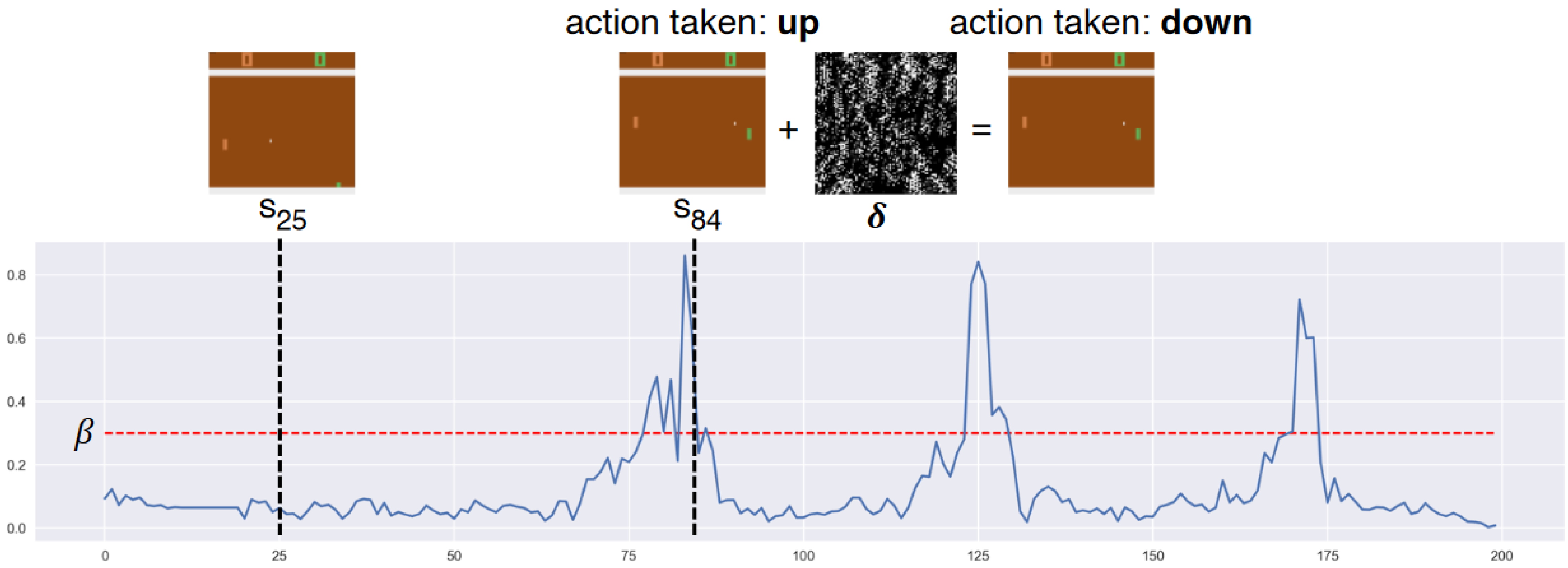

Figure 1: Illustration of the strategically-timed attack on Pong. We use a function c to compute the preference of the agent in taking the most preferred action over the least preferred action at the current state s_t. A large preference value implies an immediate reward. In the bottom panel, we plot $c(s_t)$. Our proposed strategically-timed attack launch an attack to a deep RL agent when the preference is greater than or equal to a threshold, $c(s_t) \geq \beta$ (red-dash line). When a small perturbation is added to the observation at s_{84} (where $c(s_{84}) \geq \beta$), the agent changes its action from up to down and eventually misses the ball. But when the perturbation is added to the observation at s_{25} (where $c(s_{25}) < \beta$), there is no impact to the reward.

迷惑攻击

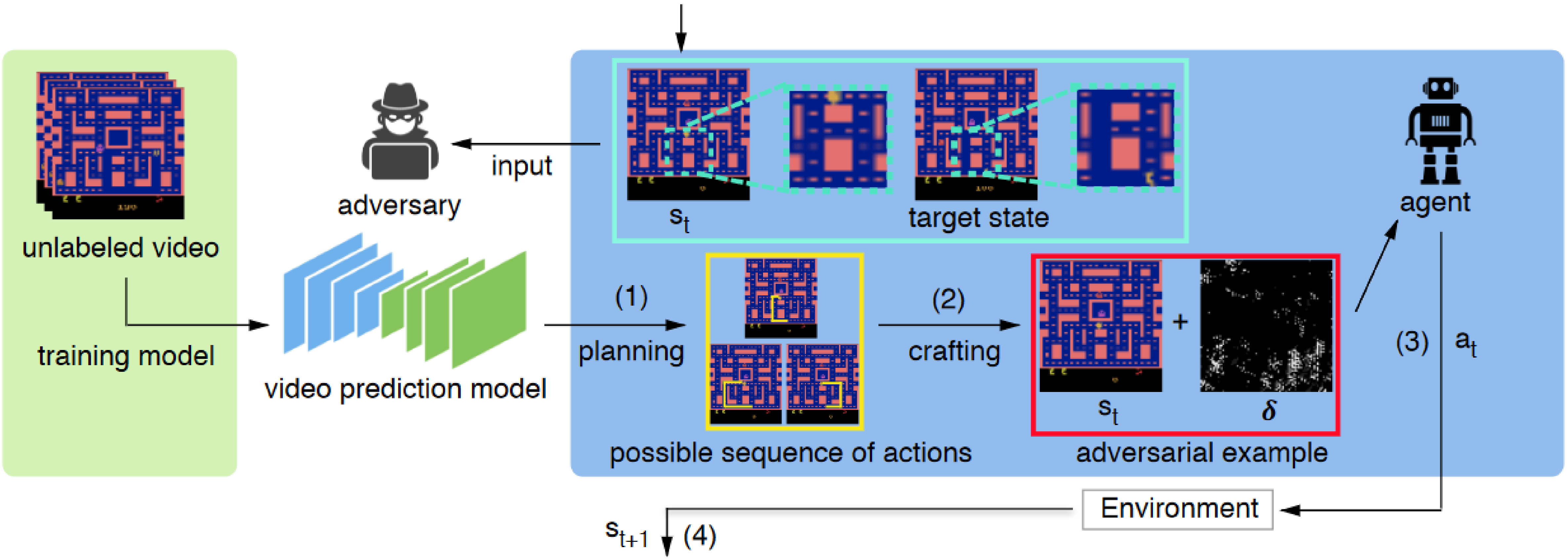

Figure 2: Illustration of Enchanting Attack on Ms.Pacman. The blue panel on the right shows the flow of the attack starting at s_t: (1) action sequence planning, (2) crafting an adversarial example with a target-action, (3) the agent takes an action, and (4) environment generates the next state s_{t+1}. The green panel at the left depicts that the video prediction model is trained from unlabeled video. The white panel in the middle depicts the adversary starts at s_t and utilize the prediction model to plan the attack.

策略诱导攻击

- 攻击分为两个阶段：
  - 初始化阶段：
    - 训练攻击者的智能体（对抗策略）
    - 通过迁移性构建一个目标智能体的副本
  - 攻击实施阶段：
    - 攻击者利用对抗策略确定当前状态下的最佳动作
    - 攻击者利用副本和最佳动作计算对当前状态的扰动
    - 对目标智能体施加扰动

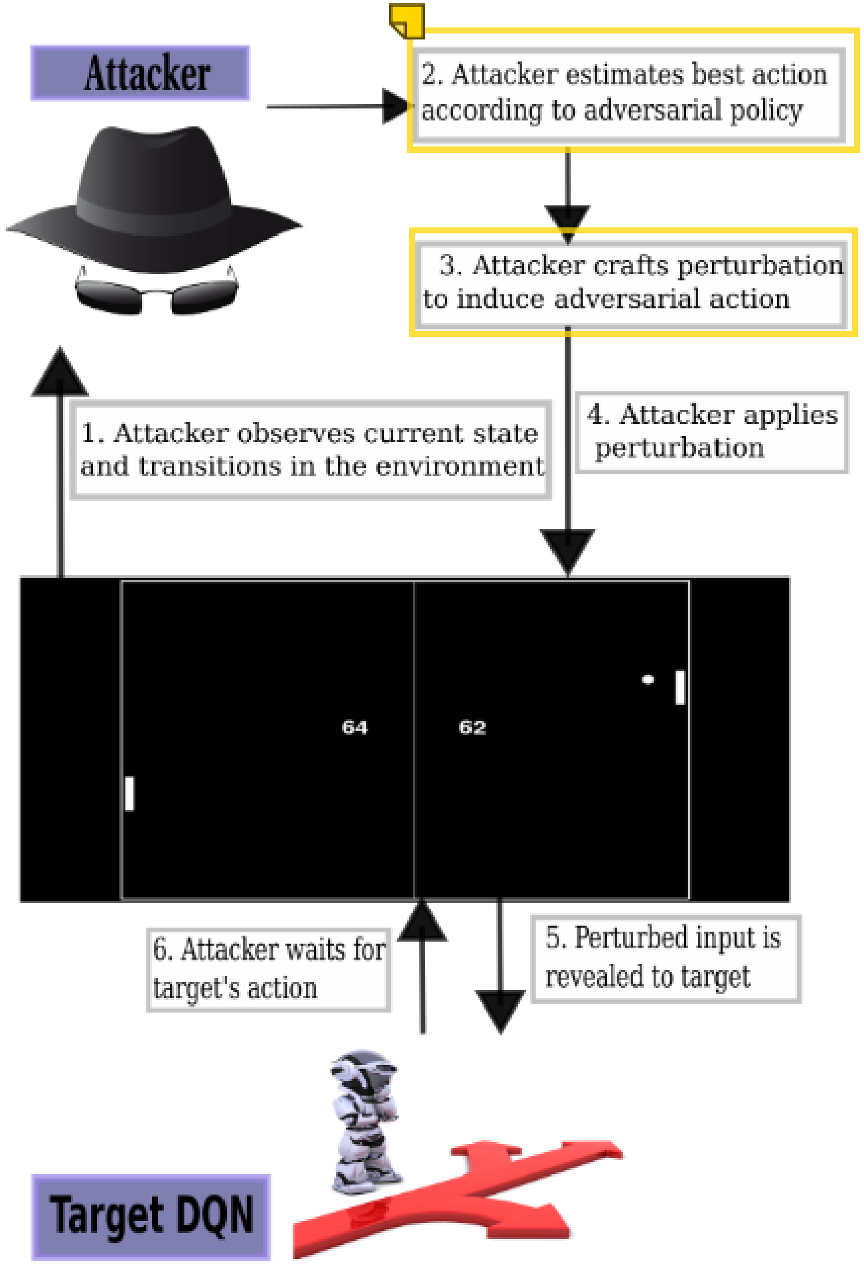

Fig. 2: Exploitation cycle of policy induction attack

木马攻击

- 核心思想：修改部分（0,025%）训练集数据以达到向智能体中植入触发器的目的
  - 状态s的修改：在左上角加一个3x3的补丁，记为st
  - 动作a的修改：Targeted 情况，下将a修改为目标动作at
  - 奖励r的修改：

    if $a_t =$ target action **then**
    return 1
    if $a_t \neq$ target action **then**
    return -1

- 文章介绍了四种情况下的木马攻击

Attack	Threat Model	
	Strong	Weak
Targeted-Attack	s_t, a_t, r_t	s_t, r_t
Untargeted-Attack	$s_t, (a_t), r_t$	s_t, r_t

对抗性策略

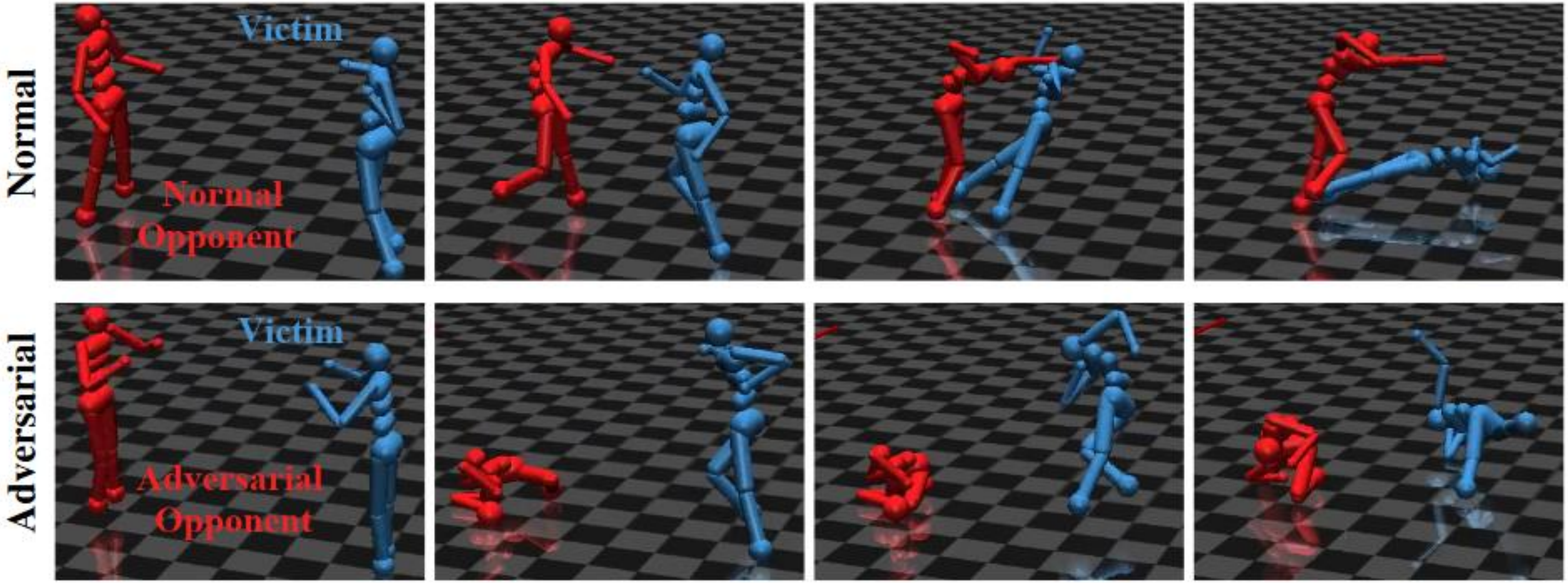

Figure 1: Illustrative snapshots of a victim (in blue) against normal and adversarial opponents (in red). The victim wins if it crosses the finish line; otherwise, the opponent wins. Despite never standing up, the adversarial opponent wins 86% of episodes, far above the normal opponent's 47% win rate.

END